Mathematics of Data Science
(Advanced topics)
Stat 946 Winter 2021

Aukosh Jagannath

Department of Statistics and Actuarial Science, University of Waterloo
Email address: a.jagannath@uwaterloo.ca

This research is funded by NSERC.

Abstract. These are the course notes for Stat 946 in Winter 2021. These notes are a compilation of scribe notes that have been modified. The scribes are: Jose Luis Avilez, Kelly Ramsay, Mahbod Majid, Liyuan Zheng

Contents

CHAPTER 1

Lecture 1: High dimensions

1. Two examples in high dimensions

1.1. "The curse of high dimensionality": Acceptance-Rejection sampling. How do I generate i.i.d. samples from the uniform measure on B_d, the unit ball in in $\mathbb{R}^d$? A classical approach from sampling from an "irregular" domain is what's called *"acceptance-rejection sampling",* one of the simplest "Monte Carlo" algorithms.

The algorithm works like this: we leverage the fact that that since $B_d \subseteq [-1,1]^d$, we can just view $Unif(B_d)$ as just $Unif([-1,1]^d)$ conditioned on $B_d(1)$. So the move is to generate $X_i \sim Unif([-1,1]^d)$ iid and then accept only those X_i which satisfy the constraint $\{||X_i||_2 \leq 1\}$. To verify that this is a good idea we can just check: if we let $\mu_n = \frac{1}{n}\sum \delta_{X_i}$ denote the empirical of the $X_i's$ then we have that

$$\mu_n(B_d(1)) \to \mu(B_d(1))$$

where μ is the uniform measure, so with enough samples I'm sampling from B_d with the right right. So if I then want to ask a statistical question like $X_i \in A$ then, if we look at the empirical measure for the accepted samples, we'd have:

$$\frac{\mu_n(B_d(1) \cap A)}{\mu_n(B_d(1))} \to \mu(A|B_d)$$

as desired. So this seems like a good move, just sample enough times and do acceptance-rejection sampling. But now suppose I actually asked you to do this in, say, $d \sim 30$. You might say, well, whatever, dimension is just a number no big deal. So you write your code and you look at your answer and you realize— "Wait...I'm rejecting *everything*!" So maybe there's a bug?

We'll let's think, whats the probability of acceptance? For a single sample its just $\mu(B_d) = \frac{Vol(B_d)}{Vol([-1,1]^d)} = \frac{Vol(B_d)}{2^d}$ Now for d even this is like

$$Vol(B_d) = \frac{\pi^{d/2}}{\Gamma(1+\frac{d}{2})} \simeq \frac{\pi^{d/2}}{\sqrt{2\pi\left(\frac{d}{2}\right)}\left(\frac{\frac{d}{2}}{e}\right)^{\frac{d}{2}}}$$

so we have something like

$$\frac{1}{d}\log \mu(V_d) \simeq -\log d + o(\log d).$$

Being a bit more careful you see that

$$\frac{Vol(B_d)}{2^d} \simeq \frac{1}{\sqrt{\pi d}}\left(\frac{\pi e}{2}\right)^d e^{-d\log d}$$

when $d \sim 15$ is like 10^{-11}, i.e., its just at the edge of what you could handle using a *double* using the IEEE 754 standard...

But this is actually kind of weird right? Draw the sphere in the cube in 3d and you see that there's basically nothing there. Then you ask yourself: what's the volume of the unit cube: 1 whats the volume of the unit sphere? Putzing around with the above you see that your only hope is when $Vol(B_d(R))$ no longer has the $d^{\frac{d}{2}}$ correction... so we'll need $R \simeq \gamma\sqrt{d}$ for some carefully chosen γ.... (something like $\gamma \sim \frac{2\sqrt{2}e}{\pi}$... but don't quote me on this....) More on this soon....

1.2. Covariance estimation. Suppose that we have iid $\mathbb{R}^d$-valued data $(X_i)_{i=1}^n$ with, say, finite k-th moment, or even say a finite MGF, that I happen to know a priori is centered and you ask me to produce an estimate of the covariance matrix. So one thing I can do is look at

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$$

and then I crack open my classic text books and say $\mathbb{E}\hat{\Sigma}_n = \Sigma$ and $\hat{\Sigma}_n's$ entries are fairly reasonable so I declare that "$\hat{\Sigma}_n$ is a great estimator because it's consistent and unbiased". Just take n large and you're good to go. But for you d is large and its expensive to produce samples so you come back to me and ask about the sample complexity: "but.... how large is n? ". For you its only reasonable to assume that I can produce like twice as many samples as the dimension.

Now you can sit down and produce an estimate... but then you realize: "Hey, I wonder what happens if you look at the spectrum of $\hat{\Sigma}_n$", so I can look at is ranked eigenvalues say λ_i^n and then look at the empirical spectral distribution $\mu_n = \frac{1}{d}\sum_{i=1}^d \delta_{\lambda_i^n}$. Suppose that my data were actually honest to goodness standard gaussians, $X_i \sim \mathcal{N}(0, I_d)$. Now if things are going to be good to go then you want this to converge to a Dirac mass at 1. But I sit there and calculate and I realize that if I take $d = \alpha n$ then

$$\mu_n \to \mu_{MP}$$

the Marchenko-Pastur law where

$$\mu_{MP} = \begin{cases} \nu_\alpha & 0 < \alpha \leq 1 \\ \nu_\alpha + \left(1 - \frac{1}{\alpha}\right)\delta_0 & \alpha > 1 \end{cases}$$

where

$$v_\alpha(dx) \propto \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{\alpha x}dx$$

and $\lambda_\pm = (1 \pm \sqrt{\alpha})^2$. In particular, this has density is very much *not* a Dirac mass and actually has its mode well below 1.

2. Concentration

One of the basic tools of our course will be concentration inequalities. Here we'll be interested in inequalities of the form

$$P(|F(X) - m| > t) \leq \alpha(t)$$

where X is an $\mathcal{X}$-valued random variable, $m \in \mathbb{R}$ and $t > 0$. For this to be a good inequality you're going to wan a a minimum that $\alpha(t) \to 0$ as $t \to \infty$ and ideally that α is (eventually) monotone in t.

Ok so what do we know about concentration. So there are some things you're probably used to

EXAMPLE 2.1. Markov type inequalities. Suppose that X is a real-valued random variable with $X \in L^p$ then you have

$$P(|X - m| > t) \leq \frac{||X - m||_p^p}{t^p}.$$

If X has an MGF then you could even go for (lets take $m = 0$ here)

$$P(|X| > t) \leq 2e^{\psi(\lambda) - \lambda t}$$

where $\psi(\lambda) = \log \mathbb{E}e^{\lambda |X|}$ is the log-Laplace transform of $|X|$. Since $\psi'' = \int (x - \int x d\mu)^2 \geq 0$ where μ is the tilted distribution, $\mu(A) = \frac{\mathbb{E}\mathbb{1}_A \exp e^{\lambda X}}{\mathbb{E}e^{\lambda X}}$, we have that ψ is convex and so if we let $\psi^*(t)$ denote its Legendre transform then we have

$$P(|X| > t) \leq 2e^{\psi^*(t)}.$$

Let's rewind to probability 101. What you're probably used to thinking about is sums of independent random variables. So here we have $S_n = \sum_{i=1}^n X_i$ where X_i are iid with, say, finite second moment. then you know

$$\frac{S_n}{n} \to \mathbb{E}X \qquad \textit{and} \qquad \frac{S_n - n\mathbb{E}X}{\sqrt{n}} \to \mathcal{N}(0, \sigma^2)$$

where the first convergence is in any mode that you want, but the second is in law. So if I think about, say $\mathbf{X} = (X_1, ..., X_n)$ as a random vector then $F(\mathbf{X}) = \frac{S_n(\mathbf{X})}{n}$ would give you something like

$$P(|F(\mathbf{X}) - \mathbb{E}X| > \epsilon) \leq \alpha(t).$$

Now using the CLT you'd see something like

$$P(|S_n - \mathbb{E}X| > \sqrt{\epsilon} n) \to P(|Z| > t)$$

where Z is said gaussian. So you might think "well, maybe I could get a bound like

$$P(|S_n - \mathbb{E}X| > \epsilon n) \leq P(|Z| > \sqrt{n}\epsilon) \leq e^{-n\epsilon^2/2}."$$

But of course that's completely unrealistic, because if that were true you would have that S_n would have light tails which won't happen if S_n is power law-type. You could try a Berry-Esseen type move but this would only give you:

LEMMA 2.2 (Berry-Esseen). *If X_i are iid mean with finite third moment then*

$$\sup_{t \in \mathbb{R}} |P(\frac{S_n}{\sqrt{n}} \geq t) - P(g \geq t)| \leq \frac{C\mathbb{E}||X||^3}{\sqrt{n}}$$

[N.b., this bound is essentially sharp: if we take t to be small and look at $P(|\frac{S_n}{\sqrt{n}}| < t)$ in the case of Bernoulli's then we'd have $P(\frac{S_n}{\sqrt{n}} = 0) \simeq \frac{1}{\sqrt{n}}$.] A better move would be to recall your elementary large deviations theory:

LEMMA 2.3 (Cramér). *Let (X_i) be iid with finite MGF and mean μ, and let $\psi(\lambda)$ denote the corresponding Cramer transform. Then if we let $I(t) = \psi^*(t)$ we have we have that*

$$-\inf_{t\in A^0} I(t) \leq \underline{\lim}_{n\to\infty} \frac{1}{n}\log P(\frac{S_n}{n} - \mu \in A) \leq \overline{\lim}\, \frac{1}{n}\log P(\frac{S_n}{n} - \mu \in A) \leq -\inf_{t\in \bar{A}} I(t).$$

What does this tell you? In plain english it says that for n large enough (and I symmetric continuous and increasing in $t > 0$)

$$P(|\frac{S_n}{n} - \mu| > \epsilon) \asymp e^{-n\min_{|t|>\epsilon} I(t)}$$

where the key point is that this is also a **lower bound.** This is pretty powerful stuff. For example, it tells you the chance of tossing a coin n times and having more than $1/2 + \epsilon$ of them being positive is like Ce^{-cn} so if n is large enough this is again something you're not going to see numerically....

Now an issue you might have with this statement is the "n large enough" part—we'll want to want to work with all $n \geq 1$ because we're going to be interested in "sample complexity" type bounds, i.e., understanding the precise trade off between n and d. But the observation is that the "hard" part of large deviations—the matching lower bound— is the part that costed us the "large enough" part. If we just focus on the corresponding upper bound, which follows from the Markov type bound from above, which gives

$$P(|\frac{S_n}{n} - \mu| > \epsilon) \leq Ce^{-nI(t)}.$$

Now this is all fine and dandy. But what if we're not working with a sums of IIDs? We're going to be interested in things like eigenvalues, empirical spectral distributions? These are not going to depend linearly on the coefficients of my matrix. Or worse: what about community detection problems? manifold learning?

CHAPTER 2

Lecture 2: High dimensional probability II

1. Hypercube Example

- The unit cube game
- Radius of unit volume sphere

2. Gaussian random variables

Our starting point for discussing concentration will be to understand Gaussian random variables.

If $Z \sim \mathcal{N}(0,1)$ then its tail is of he form

$$P(|Z| > t) = 2\int_t^\infty e^{-\frac{s^2}{2}} ds$$

Now we recall Mill's inequality which says that

$$\frac{\lambda}{\lambda^2+1} e^{-\lambda^2/2} \leq P(|Z| > \lambda) \leq \frac{1}{\lambda} e^{-\lambda^2/2}$$

So in particular we have concentration with a tail like $e^{-\lambda^2/2}$.

What happens now when we think about multi variate gaussians? Recall that a multivariate gaussian is a random variable $X \in \mathbb{R}^d$ denoted by $X \sim \mathcal{N}(\mu, \Sigma)$ which has mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$ (in fact it's positive semidefinite).

More generally you could consider a Gaussian process which is a family of gaussians $(X_t)_{t\in T}$ where T is some set. Examples of gaussian processes:

EXAMPLE 2.1. A gaussian random vector Here we have $T = [n]$

EXAMPLE 2.2. The isonormal process on $\mathbb{R}^d$,

$$g(v) = \langle g', v \rangle ,$$

where g' is a standard gaussian vector.

EXAMPLE 2.3. Brownian motion

The key thing to keep in mind is that Gaussian processes are characterized by their means and their covariance kernels

$$\begin{aligned} \mu(t) &= \mathbb{E}X_t \\ C(t,s) &= \mathbb{E}\left[(X_t - \mu(t))\,(X_s - \mu(s))\right] \end{aligned}$$

Let's not get too exotic today though, and let's just focus today on the simple case where of standard Gaussian's on $\mathbb{R}^d$.

3. Concentration of the norm of a standard Gaussian

So now we have a question: where does a gaussian random variable like to live?

THEOREM 3.1. *Let $X \sim \mathcal{N}(0, I_n)$ then for for some $t_0 > 0$ universal, $0 < t < c$ we have*

$$P(|||X||_2 - \sqrt{n}| > \sqrt{n}t) \leq 2e^{-cnt^2/2}$$

PROOF. [first proof] The idea is as follows. let's look at

$$||X||_2^2 = \sum X_i^2$$

now we have that for $\lambda < \frac{1}{2}$,

$$\mathbb{E}e^{\lambda X_1^2} = \sqrt{\frac{1}{1-2\lambda}}$$

so

$$\begin{aligned} P(\sum X_i^2 > (1+t)n) &\leq e^{-\lambda(1+t)n}\mathbb{E}\exp\left(\lambda \sum X_i^2\right) \\ &\leq \exp\left[n\left(-\lambda(1+t) - \frac{1}{2}\log(1-2\lambda))\right)\right] \\ &\leq \exp\left[-\frac{n}{2}t^2\left(\frac{1}{2} - \frac{t}{3}\right)\right] \end{aligned}$$

where to go from line 2 to line 3 we optimized with $\lambda_* = \frac{t}{2(1+t)} < \frac{1}{2}$ and Taylor expanded $t - \log(1+t)$ to third order. Thus if $t < \frac{3}{2}$ we're good to go. By symmetry we get the reverse bound. □

- discussion about thin shells

4. The Johnson-Lindenstrauss lemma

One application of the above is for low distortion embeddings.

DEFINITION 4.1. We say that a map metric spaces $f : X \to Y$ is an $\epsilon-$quasi isometry if

$$(1-\epsilon)d_X(x,y) \leq d_Y(f(x), f(y)) \leq (1+\epsilon)d_X(x,y)$$

THEOREM 4.2. *Let $0 < \epsilon < \frac{1}{2}$ and $Q \in \mathbb{R}^d$ have $|Q| = n$ and then there is some C universal such that for $k \leq \frac{C\log n}{\epsilon^2}$ then there is a Lipschitz map $f : \mathbb{R}^d \to \mathbb{R}^k$ which is n ϵ -quasi isometry of on Q.*

We'll show here that in fact its really easy to construct such a map. Furthermore the mapping is data and dimension agnostic!

THEOREM 4.3. *under the same conditions as above, if A is a $k \times d$ matrix with iid $N(0,1)$ entries then we have that $f(x) = \frac{1}{\sqrt{k}}Ax$ is an $\epsilon-$quasi isometry with probability $1 - Cn^2\exp(-ck\epsilon^2)$ for $\epsilon < \epsilon_0$ for some $\epsilon_0 = t_0$ from above.*

PROOF. By a union bound we have that

$$P(\exists x, y : \text{not a good embedding }) = n^2 \max_{\substack{u=x-y \\ x,y\in Q}} P(\frac{||Au||}{\sqrt{k}||u||} \in [1-\epsilon, 1+\epsilon]^c)$$

Now just note that $\frac{Au}{|u|} \sim N(0, I_k)$ so we're done by the above, which shows that the preceding is bounded above by $Cn^2 \exp(-ck\epsilon^2)$ which is small provided $k \geq \frac{K}{\epsilon^2} \log n$ for some large enough K $\square$

5. Non-linear functions?

For example, here's a question. Recall the Gaussian orthogonal ensemble. That is a random matrix $A = (a_{ij})$ whose entries on the upper triangle are iid $Var(a_{ij}) = N(0, \frac{1+\delta_{ij}}{N})$ and on the lower triangle its the symmetric reflection i.e. $a_{ij} = a_{ji}$. This class of random matrices appears allover the place in statistics, data science, and mathematical physics. One question you might ask yourself is what can you say about the top-eigenvalue of A, $\lambda_1(A)$. In particular can you show concentration of the form:

$$P(|\lambda_1(A) - \mathbb{E}\lambda_1(A)| \geq t) \leq \alpha(t, N)$$

in a similar vein, can you bound the *operator norm* of the matrix (this is sometimes called the $(\ell_2 \to \ell_2)-$norm),

$$||A||_{op} = \sup_{||u||_2=1} |\langle u, Au \rangle|$$

that is can you show something like

$$P(|||A||_{op} - \mathbb{E}||A||_{op}| \geq t) \leq \alpha(t, N)?$$

This is in and of itself not completely obvious. So the idea is to use the following meta theorem

THEOREM. *(Meta-theorem) A Lipschitz function of a "reasonable" random variable in high-dimensions concentrates about its mean.*

Let's start with the following.

THEOREM 5.1. *There are universal constants $C, c > 0$ such that the following holds. Let $X \sim N(0, I_d)$ for any $d \geq 1$ and let $F : \mathbb{R}^d \to \mathbb{R}$ be K-Lipschitz with respect to the euclidean distance $|f(x) - f(y)| \leq K||x - y||$, then $F(X)$ has a mean and for any $t \geq 0$ we have*

$$P(|F(X) - \mathbb{E}F(X)| \geq t) \leq C \exp(-c\frac{t^2}{K^2}).$$

PROOF. That there is a mean is clear. (Why?) Lets turn to the rest. We'll also focus on the case F is smooth.

The idea we'll use here is whats called the *smart path method.* Let X, Y denote two independent copies of this process, and for $0 \leq t \leq 1$ let $X_t = \cos(t)X + \sin(t)Y$. Notice that X_t is differentiable with $\dot{X}_t = -\sin(t)X + \cos(t)Y$. In particular, for each t, X_t and $\dot{X}_t$

are both mean zero and are identically distributed. Now we compute

$$\begin{aligned}
\mathbb{E}\exp\left(\lambda(F(X)-\mathbb{E}F(X))\right) &\leq \mathbb{E}\exp\left(\lambda(F(X)-F(Y)\right) \\
&= \mathbb{E}\exp\left(\lambda\int_0^{\pi/2}\left\langle \nabla F(X_t), \dot{X}_t\right\rangle dt\right) \\
&\leq \fint_0^{\pi/2}\mathbb{E}\exp\left(\lambda\frac{\pi}{2}\left\langle \nabla F(X_t), \dot{X}_t\right\rangle\right) dt \\
&= \fint_0^{\pi/2}\mathbb{E}\left[\exp\left(\frac{\lambda^2\pi^2}{4}||\nabla F(X_t)||^2\right)\right] dt \\
&\leq e^{\frac{1}{2}K^2\frac{\lambda^2\pi^2}{4}}
\end{aligned}$$

where in the first and third line we used Jensen. Thus for $\lambda, t \geq 0$ we have

$$P(|F(X)-\mathbb{E}F(X)|\geq t) \leq 2\exp\left(-\lambda t + \frac{\pi^2}{4}K^2\lambda^2\right)$$

Recall now that $\frac{a}{2}t^2$ has Legendre transform $\sup_{\lambda\geq 0}\lambda t - \frac{a}{2}\lambda^2 = \frac{t^2}{2a}$, so that by minimizing the above we have

$$P(|F(X)-\mathbb{E}F(X)|\geq t) \leq 2\exp\left(-\frac{2}{\pi^2}\left(\frac{t}{K}\right)^2\right)$$

as desired. □

CHAPTER 3

Lecture 3: Gaussian Concentration
Scribe: J. Avilez

For this lecture, we will be working with the Gaussian Orthogonal Ensemble, so we recall a definition.

DEFINITION 0.1. A matrix $A = (a_{ij}) \in M_N(\mathbb{R})$ is said to belong to the Gaussian Orthogonal Ensemble (GOE) if it is symmetric and its upper triangular entries are given by $a_{ij} \sim N\left(0, \frac{1+\delta_{ij}}{N}\right)$, where the entries in the upper triangular part are independent. If A is a GOE matrix of dimension $N \times N$, we write $A \sim GOE(N)$

In this lecture we will be concerned with concentration for three objects associated to a GOE matrix:

(1) Its top eigenvalue;
(2) Its empirical spectral measure;
(3) Its operator norm.

1. Top eigenvalue

To discuss concentration for the top eigenvalue of a GOE matrix we will need a theorem by Borell. No, that is not a typo.

THEOREM 1.0.1. ***(Borell's inequality)*** *Let T be a compact metric space and $\{X_t\}_{t\in T}$ a centred Gaussian process. Suppose that:*

(1) *The covariance kernel $K : T \times T \to \mathbb{R}$ given by $K(t,s) = \mathbb{E}X_tX_s$ is continuous.*
(2) $\sup_{t\in T} X_t < \infty$ *almost surely*

Then $\mathbb{E}\sup_{t\in T} X_t < \infty$ *and*

$$P\left(\left|\sup_{s\in T} X_s - \mathbb{E}\sup_{s\in T} X_s\right| > t\right) \leq 2e^{-\frac{t^2}{2\sigma_T^2}}$$

Where $\sigma_T^2 = \sup_{t\in T} \mathbb{E}X_t^2$

PROOF. Recall that if f is K-Lipschitz, then for a vector of independent standard Gaussians, $P\left(|f(Z) - \mathbb{E}f(Z)| > t\right) \leq 2e^{-\frac{t^2}{2K^2}}$.

In the case T is finite, we observe that the maximum over T is a 1-Lipschitz map, so the result follows from our metatheorem.

It's a good thing that compactness is just the generalisation of being finite. To lift the argument up the general compact metric T, take a sequence of finite sets $T_1 \subset T_2 \subset T_3 \subset \ldots$ that increases to a dense subset D of T. Then, argue via monotone convergence

that the required quantities converge adequately. Finally, show that assumption 2 will yield $\mathbb{E}\sup_{t\in T} X_t < \infty$[1]. □

We can apply Borell's inequality to discuss the top eigenvalue of a GOE matrix.

EXAMPLE 1.1. Suppose $A \sim GOE(N)$. Since A is Hermitian, we may write the top eigenvalue by $\lambda_N(A) = \sup_{||v||_2=1}\langle v, Av\rangle$. For a unit vector v, we will often call the quantity $R_A(v) = \langle v, Av\rangle$ the *Rayleigh quotient* of A at v. Note that $R_A(v)$ is a linear combination of the entries of A, so since A is a GOE matrix, $\{R_A(v)\}_{v\in S^{N-1}}$ is a centred Gaussian process. The covariance kernel can be computed to be:

$$K(u,v) = \mathbb{E}R(u)R(v) = \mathbb{E}\langle u, Au\rangle\langle v, Av\rangle = \frac{1}{N}\left(\langle u,v\rangle^2 + \langle u,v\rangle\right)$$

Note that the covariance kernel is continuous. We may now apply Borell's inequality to get:

$$\begin{aligned} P\left(|\lambda_N(A) - \mathbb{E}\lambda_N(A)| > t\right) &= P\left(|\sup_{||v||_2=1}\langle v, Av\rangle - \mathbb{E}\sup_{||v||_2=1}\langle v, Av\rangle| > t\right) \\ &\leq 2e^{-N\frac{t^2}{2}} \qquad \text{(Borell)} \end{aligned}$$

2. Concentration for empirical spectral measure

A random matrix will have a random spectrum; to it, we associate an average:

DEFINITION 2.1. Let A be an $N \times N$ random matrix. Its **empirical spectral measure** is the random measure given by:

$$\mu_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{\lambda_i}$$

Where $\{\lambda_1, \ldots, \lambda_N\}$ is the spectrum of A.

Next week we will prove the following surprising theorem:

THEOREM 2.1.1. *Let $A_N \sim GOE(N)$. The empirical spectral measure satisfies the following asymptotics: $\mu_N \to \sigma_{sc}(dx)$ weakly almost surely, where $\sigma_{sc}(dx) = \sqrt{4-x^2}dx$ is the semi-circle law.*

Before we get there, we will need results pertaining to the concentration of random measures. To that end, we need a suitable metric for the space of probabilities of $\mathbb{R}$.

DEFINITION 2.2. Let $\mu, \nu \in \text{Prob}(\mathbb{R})$ be Radon probability measures. The **Wasserstein distance** between μ and ν is the quantity:

$$d_W(\mu,\nu) = \sup_{f\in\mathcal{L}}|\int f d\mu - \int f d\nu|$$

Where $\mathcal{L} = \{f \in C_b(\mathbb{R}) : ||f||_\infty \leq 1 \text{ and } ||f||_{Lip} \leq 1\}$

REMARK 2.3. The Wasserstein distance is a metric on the space of Radon probability measures Prob($\mathbb{R}$). Morally, the Wasserstein distance measures the optimal cost of transporting the pile of stuff μ to ν.

[1]For full details see p.56-57 of Random Fields and Geometry by Adler and Taylor.

The following concentration inequality is attributed to several people.

THEOREM 2.3.1. *(Guionnet, Ben Arous, Guionnet, Zeitouni) For any μ, $N \geq 1$:*

$$P(d_W(\mu_N, \mu) \geq \mathbb{E}d_W(\mu_N, \mu) + \delta) \leq 2e^{-cN^2\delta^2}$$

We need to use the Hoffman-Wielandt inequality: given two $N \times N$ real symmetric matrices A, B, with sorted eigenvalues $\lambda_1(A) \leq \ldots \leq \lambda_N(A)$, and $\lambda_1(B) \leq \ldots \leq \lambda_N(B)$, the following inequality holds:

$$\sqrt{\sum_{i=1}^{N}(\lambda_i(A) - \lambda_i(B))^2} \leq ||A - B||_F = \text{tr}\left((A-B)(A-B)^T\right)$$

Where $||\cdot||_F$ denotes the Frobenius norm. Two corollaries of this finding are:

(1) If $g : \mathbb{R}^N \to \mathbb{R}$ is 1-Lipschitz, then $g(\lambda_1, \ldots, \lambda_N)$ (where we think of g as being a function of A) is 1-Lipschitz for $||\cdot||_F$;
(2) If $f : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz, then the map $A \mapsto \text{tr}(f(A))$ is 1-Lipschitz[2].

We may now proceed with the proof of concentration.

PROOF. (Of Theorem 2.3.1) Let $F(A) = d_W(\mu_N, \mu) = \sup_{f \in \mathcal{L}} |\int f d\mu_N - \int f d\mu|$. Observe that F is $\frac{1}{N}$-Lipschitz. Furthermore, the map $A \mapsto \int f d\mu_N \frac{1}{N}\sum_{i=1}^{N} f(\lambda_i) = \frac{1}{N}\text{tr}(f(A))$ is $\frac{1}{N}$-Lipschitz for all $f \in \mathcal{L}$.

Using standard Gaussian concentration, we are done. □

3. Operator norms

Recall from Hilbert space theory that for symmetric matrices $||A||_{op} = \sup_{||x||=1} |\langle x \rangle Ax|$.

We may use Borell's inequality, or other results from Gaussian concentration, to show that for a GOE matrix A we have $P\left(|||A||_{op} - \mathbb{E}||A||_{op}| > t\right) \leq e^{-cNt^2}$.

This inequality tells us that our random variable $||A||_{op}$ lives around some constant. However, this inequality tells us nothing about what that constant is!

Using some crude bounds we may observe:

$$||A||_{op} \leq ||A||_F = \sqrt{\sum_{i,j=1}^{N} a_{ij}^2} = \sqrt{\sum_{i,j=1}^{N} \underbrace{||Ae_i||^2}_{\leq ||A||_{op}^2}} \leq \sqrt{n}||A||_{op}$$

Then, our guess based on these crude bounds is that $||A||_F^2 \approx \sum a_{ij}^2 \approx \frac{1}{N}\sum g_{ij}^2 \sim \frac{N^2}{N} = N$. This tells us that the Frobenius norm scales like $\sqrt{N}$. A more intelligent approximation may be to use the semi-circle asymptotics for GOE (Theorem 2.1.1), we may guess that $||A||_{op} \sim 2$. These two approaches have a huge discrepancy for random matrices! Perhaps an initial explanation for this discrepancy is that $||g||_2 \sim \sqrt{d}$ and $||g||_\infty \sim \sqrt{\log d}$

Before we formalise this discussion into a result, we recall a definition from metric space theory.

DEFINITION 3.1. Let $\epsilon > 0$ and let X be a metric space. An ϵ-net N_ϵ is a set such that for all $y \in X$ there exists $x \in N_\epsilon$ such that $d(x, y) \leq \epsilon$

[2]See appendix on spectral theory to see what $f(A)$ means.

A consequence of the triangle inequality is that given an ϵ-net of S^{n-1} for $0 < \epsilon < 0.5$, we have (check!):

$$||A||_{op} \leq \frac{1}{1-2\epsilon} \sup_{x \in N_\epsilon} |\langle x, Ax \rangle|$$

So if we set $\epsilon = 1/4$, we get $||A||_{op} \leq 2 \sup_{x \in N_\epsilon} |\langle x \rangle Ax|$.

We now show that $||A||_{op}$ is $O(1)$.

THEOREM 3.1.1. *For $A \sim GOE(N)$ we have $P(||A||_{op} > C(1+t)) \leq e^{-ct^2 N}$.*

PROOF. Take a $\frac{1}{4}$-net N_ϵ of the unit sphere. Then,

$$\begin{aligned}
P(||A||_{op} > u) &\leq P\left(\sup_{x \in N_{1/4}} |\langle x, Ax \rangle| > 2u \right) && \text{(net bound)} \\
&\leq \underbrace{|N_{1/4}|}_{c^N} \; P(|\underbrace{\langle x, Ax \rangle}_{N(0,1/N)}| > 2u) && \text{(union bound)} \\
&\leq \exp\left(N \left(\log c - \frac{(2u)^2}{2} \right) \right) \\
&\leq e^{-cNu^2}
\end{aligned}$$

Where the step from line 2 to 3 depends on an exponential bound on the size of an ϵ-net of the unit sphere and the last inequality holds for u sufficiently large. $\square$

Let's address the elephant in the room: the size of an ϵ-net of S^{N-1}. First we observe that if two points in the sphere are close, then their inner product is close to 1. A few questions:

- What is the maximum number of points in S^{N-1} have $\langle x_i, x_j \rangle = 0$. N, duh!
- How many points in S^{N-1} have $|\langle x_i, x_j \rangle| \leq \epsilon$. It's about $e^{nc(\epsilon)}$

Let's see the second point. Say $x_1 \sim Unif(S^{N-1})$. Then:

$$P(x_1 > \epsilon) = P\left(\frac{g_1}{||g||} > \epsilon \right) \approx P(g_1 > \epsilon\sqrt{n}) \approx e^{-n\epsilon^2}$$

Morally, this means that if we take a spherical cap that does not contain the equator, then it has exponentially small mass.

CHAPTER 4

Lecture 4: Isoperimetry, Log-Sobolev, and Poincaré
Scribe: J. Avilez

In this lecture we will discuss tools and techniques to describe concentration in more complicated settings. This lecture will focus on three points: (i) isoperimetric inequalities, (ii) Poincaré inequalities, and (iii) log-Sobolev inequalities.

1. Isoperimetric inequalities

The basic question in isoperimetry is: for a given amount of volume, what shape maximises volume over surface area? A classic result is that in $\mathbb{R}^n$ a sphere has the smallest area per given volume. Ultimately, we wish to relate isoperimetry with concentration.

For the remainder of the section we will work on a metric measure space; that is, a triple (X, d, μ) where (X, d) is a metric space and μ is a Borel probability measure on (X, d).

Definition 1.1. The **concentration function** on (X, d, μ) is:

$$\alpha_{(X,d,\mu)}(r) = \sup_{\substack{A \subseteq X \\ \mu(A) \geq 1/2}} \mu(A_r^c)$$

Where $A_r = \{x \in X : d(x, A) \leq r\}$.

The value $\frac{1}{2}$ is not some magical constant; it connects naturally with concentration, as it comes from answering questions about medians. For instance, if F is K-Lipschitz and m_F is a median for F, we have (exercise):

$$\mu(\{F \geq m_F + r\}) \leq \alpha\left(\frac{r}{K}\right)$$

Here are two special cases:

- Subgaussian: $\alpha(r) \leq Ce^{-cr^2}$
- Subexponential $\alpha(r) \leq Ce^{-cr}$

A further tool we will wield in connecting isoperimetry and concentration is named after a famous geometer.

Definition 1.2. The **Minkowski content** of a Borel set $A \subset X$ is $\mu^+(A) = \lim_{r \to 0} \frac{\mu(A_r) - \mu(A)}{r}$.

Morally, the Minkowski content is the derivative of the measure μ with respect to r-blow ups of sets.

We shall say that an isopetrimetric inequality holds with $I : [0, \infty) \to \mathbb{R}_+$ if $\mu^+(A) \geq I(\mu(A))$.

Theorem 1.2.1. *Let v be monotone and differentiable with $I \geq v$. Then $\alpha(r) \leq 1 - v'\left(v^{-1}(1/2) + r\right)$*

EXAMPLE 1.3. In S^{N-1}, a hyperspherical cap is the optimal set when it comes to isoperimetry with respect to Lebesgue measure. This can be shown by calculating the volume of a hyperspherical cap (calculus exercise):

$$v(r) = \frac{\int_0^r \sin^{n-1}(x)dx}{s_n}$$

Using more calculus we can get the result $1 - v'\left(v^{-1}(1/2) + r\right) \leq e^{-\frac{(n-1)}{2}r^2}$

EXAMPLE 1.4. A more exotic setting where we have isoperimetry is infinite-dimesional Gaussian space. That is, we consider the measure space $(\mathbb{R}^\infty, \gamma_\infty)$. (N.B. I suspect the metric here is $d(x,y) = \sum_{k=1}^\infty \frac{|x_k - y_k|}{1+|x_k-y_k|}$) In this case $I(r) = \Phi(r)$.

We would expect sub-Gaussian concentration in Gaussian space, and indeed we do attain it:

$$\alpha(r) \leq 1 - \Phi\left(\Phi^{-1}(1/2) + r\right) \leq Ce^{-\frac{r^2}{2}}$$

In this case, a half-space is the optimal set for isoperimetry. Proving this is not easy.

Recall that when we proved Gaussian concentration in $(\mathbb{R}^n, \gamma_n)$, where $d\gamma_n = \frac{e^{-||x||_2^2/2}}{(2\pi)^{n/2}}dx$ we appealed to the smart path method. More precisely, given two independent standard Gaussians, Z, Z', we let $Z_t = \sqrt{1-t}Z + \sqrt{t}Z'$ be the so-called "smart path". We may think of Z_t as an invariant flow for γ_n.

The folks in computer science like to define the define a related object, the noise operator, which is the expectation of the smart path that evolves from x: $T_t f(x) = \mathbb{E}f(\sqrt{1-t}x + \sqrt{t}Z')$. This can be used to study more Gaussian concentration.

Now, let's use some tools from stochastic analysis to take this idea further. If we make the exponential time change $t = e^{-s}$, so that $X_s = Z_{e^{-s}}$. We may observe that X_s is an Ornstein-Uhlenbeck process. In particular, it satisfies the stochastic differential equation $dX_t = \sqrt{2}dB_t - X_t dt$ with initial condition $X_0 = x$. Of note: X_s is a reversible Markov process whose invariant measure is γ_n. In this setting, the object we should study is the Markov semi-group of X_s; let us take a detour into this subject.

2. Detour: Markov semi-groups

DEFINITION 2.1. Let $B = C(\Omega)$ be the Banach space of continuous real-valued functions on some complete separable metric space Ω, furnished with the uniform norm. A family $(P_t)_{t\geq 0}$ of linear operators on B is called a Markov semi-group if it satisfies the following properties:

(1) $P_0 = I$;
(2) The map $t \mapsto P_t$ is continuous in the sense that for all $f \in B$, $t \mapsto P_t f$ is a continuous map from $\mathbb{R}_+$ to B;
(3) $P_{t+s} = P_t P_s$;
(4) The constant function $f(x) = 1$ is a fixed point for P_t for all $t > 0$.
(5) P_t is a positive operator: $f \geq 0$ implies that $P_t f \geq 0$

REMARK 2.2. Properties 1-3 are the semi-group properties; properties 4-5 are the Markov properties. In particular, the Markov properties imply that P_t is contractive for all $t > 0$; i.e. $||P_t f|| \leq ||f||$.

EXAMPLE 2.3. (d-dimensional Brownian motion) Let $f \in C(\mathbb{R}^d)$ and $\omega \in \mathbb{R}^d$. Define the linear operators $(P_t)_{t\geq 0}$ via:

$$P_t f(\omega) = \frac{1}{(2\pi t)^{d/2}} \int f(y) e^{-\frac{||\omega - y||_2^2}{2t}} dy$$

Then $(P_t)_{t\geq 0}$ is a Markov semi-group on the space $C_{ub}(\mathbb{R}^d)$, the space of uniformly continuous bounded functions on $\mathbb{R}^d$.

Properties 1, 2, 4, and 5 are obvious. Property 3 requires a bit more work; we leave the details to the reader.

EXAMPLE 2.4. (Brownian motion on the circle) Let $\Omega = S^1$. From Fourier analysis, any $f \in C(\Omega)$ can be represented by its expansion on the Fourier basis

$$f(x) = \sum_{n=-\infty}^{\infty} a_n e^{inx}$$

Define the family of operators $(P_t)_{t\geq 0}$ by:

$$P_t f(x) = \sum_{n=-\infty}^{\infty} e^{-tn^2} a_n e^{inx}$$

Then $(P_t)_{t\geq 0}$ is a Markov semigroup on $C(S^1)$.

The only property that is not obvious (to me) is positivity. I leave the details to the reader.

EXAMPLE 2.5. (Ornstein-Uhlenbeck Process) Set $\Omega = \mathbb{R}$ and define the operator family $(P_t)_{t\geq 0}$ on $C_b(\mathbb{R})$ by:

$$P_t f(x) = \int f\left(e^{-t}x + \sqrt{1 - e^{-2t}}y\right) \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy$$

Then this defines a Markov semi-group.

An object associated to a semi-group is its infinitesimal generator.

DEFINITION 2.6. The **infinitesimal generator** of a semi-group $(P_t)_{t\geq 0}$ is defined by

$$Lf = \lim_{t \downarrow 0} \frac{1}{t}(P_t - I)f$$

for any function $f \in C_b(\Omega)$ on which the limit is well-defined. We denote this domain by $D(L)$.

EXAMPLE 2.7. For our three examples above we have:

- For 1-dimensional Brownian motion $Lf = \frac{1}{2}f''$ and $D(L) = C^2(\mathbb{R})$ with bounded first and second derivative.
- For Brownian motion on the circle the result is the same as above.
- For OU processes, we have $Lf(x) = f''f(x) - xf'(x)$, with $D(L) = C^2(\mathbb{R})$.

3. End of detour: back to smart paths

Having observed that the Markov semi-group of the Ornstein-Uhlenbeck process is $P_t f(x) = \mathbb{E} f(X_t) \,|\, X_0 = x$, we observe that the noise operator and this Markov semi-group are related to each other by an exponential time change. That is $T_t = P_{e^{-t}}$.

This is great, because we can now connect concentration and stochastic analysis, which shall become useful later in the course. A stochastic process we will care about is one that arises from mathematical physics. First, let's set the stage.

Consider the Borel measure space $(\mathbb{R}^n, \mu_u)$ where $\mu_u = \frac{e^{-u(x)}}{z} dx$ be a Gibbs measure with potential u (where z is a normalising constant). Note that if $u(x) = \frac{||x||^2}{2}$ we recover Gaussian measure.

Definition 3.1. A **Langevin process** with potential u is the stochastic process that solves $dX_t = \sqrt{2} dB_t - \nabla u(X_t) dt$ with initial data $X_0 = x$.

Ultimately we want to connect the ergodic theory of X_t with concentration for the Gibbs measure μ_u.

To do so, we work in the general setting $(\mathcal{X}, \mu)$ where $\mathcal{X}$ is a state-space with μ a measure, where X_t is a Markov process that is time-homogenous, reversible/symmetric, and μ is the invariant measure for X_t. In this setting, we can always concoct a Markov semi-group $(P_t)_{t \geq 0}$ by letting $P_t f(x) = \mathbb{E} f(X_t) \,|\, X_0 = x$. The final ingredient we will need is a certain inner product.

Definition 3.2. The Dirichlet form associated with the infinitesimal generator L is the map $\epsilon : L^2(\mu)^2 \to \mathbb{R}$ given by:

$$\mathcal{E}(f, g) = \langle f, -Lg \rangle_{L^2(\mu)} = \int f(-Lg) d\mu$$

We may ask where the Dirichlet form is well-defined, so we can just give that space a name: $H_1 = \{L^2(\mu) : \mathcal{E}(f, f) < \infty\}$.

Now that we have talked about Markov semi-groups in the context of stochastic analysis, let us introduce an extra infinitesimal generator.

Example 3.3. The infinitesimal generator of the Markov semi-group induced by the Langevin dynamics with potential u is:

$$L = \Delta - \langle \nabla u, \nabla \cdot \rangle$$

The Dirichlet form associated to Langevin dynamics has a nice form that arises from integration by parts:

$$\mathcal{E}(f, g) = \langle f, -Lg \rangle = \langle \nabla f, \nabla g \rangle$$

4. Poincaré Inequality and Spectral Gaps

Consider the diagonalisable operators P_t defined by $P_t f = e^{tL} f = \sum e^{-\lambda_k t} \langle f, u_k \rangle u_k$. A few observations about the spectrum of $-L$ are that:

(1) $\lambda_0(-L) = 0$
(2) $\lambda_1(-L) = \sup_{f \in H_1} \frac{\mathcal{E}(f,f)}{\mathrm{Var}(f)}$
(3) $\lambda_1(-L) > \gamma > 0$

Altogether, this implies that $\text{Var} f \leq \frac{1}{\lambda_1}\mathcal{E}(f,f)$. This motivates the following definition.

DEFINITION 4.1. The metric measure space (X, d, μ) has a Poincaré inequality with constant $\gamma > 0$ if $\text{Var} f \leq \frac{1}{\gamma}\mathcal{E}(f,f)$ for all $f \in H_1$.

Why does this matter? Let's consider the λ_1-eigenvector. Then we get $P_t u_1 = e^{-\lambda_1 t} u_1$, which governs the rate at which a process converges to equilibrium. We can use this to recover concentration.

THEOREM 4.1.1. *(Poincaré implies concentration) Suppose (X, μ, L) has a Poincaré inequality with constant $\gamma > 0$. For all 1-Lipschitz functions F, we have sub-exponential concentration:*

$$P\left(F > \int F d\mu + r\right) \leq e^{-\frac{\gamma}{3}r}$$

We won't prove this theorem, but we note that the idea is to go via Cheeger's inequality.

EXAMPLE 4.2. In $(\mathbb{R}^n, \gamma^n)$ we get sub-exponential concentration with $\lambda_1 = 1$. This is not great, as we expect sub-Gaussian concentration.

EXAMPLE 4.3. Let M be a compact boundaryless manifold, such as $S^n, SO(n), \ldots$, then $\lambda_1(-\Delta) > 0$, which gives us subexponential concentration.

5. Log-Sobolev inequalities

The setting here is some submanifold $\mathcal{X} = M$ with Gibbs measure $\mu = \mu_u$. In this setting we once again have $\mathcal{E}(f,g) = \langle \nabla f, \nabla g \rangle$.

DEFINITION 5.1. The space (M, μ) has a log-Sobolev inequality with constant c if:

$$\int f^2 \underbrace{\log\left(\frac{f^2}{\int f^2 d\mu}\right)}_{(\star)} d\mu \leq 2c \int |\nabla f|^2 d\mu$$

Note that if we set $\rho = f^2 d\mu$ then the section in $(\star)$ is the relative entropy of ρ with respect to μ, i.e. $H(\rho|\mu)$.

As has been a theme throughout this lecture, a log-Sobolev inequality yields concentration.

THEOREM 5.1.1. *If (M, μ) has a log-Sobolev inequality with constant c then for all 1-Lipschitz functions f we have:*

$$P\left(f > \int f d\mu + r\right) \leq e^{-\frac{r^2}{8c}}$$

PROOF. (Sketch) Take the log-MGF $\Lambda(\lambda) = \log \mathbb{E} e^{\lambda f(x)}$. Show that $\frac{d}{d\lambda}\left(\frac{\Lambda(\lambda)}{\lambda}\right) \leq$. Use Grönwall's inequality to conclude that $\mathbb{E} e^{\lambda f} \leq e^{c\lambda^2}$. □

Log-Sobolev inequalities look pretty crazy, so we wonder how to prove a space has a log-Sobolev inequality.

THEOREM 5.1.2. *(Bakry-Emery Condition) Let $c > 0$. Suppose (M, μ_u) satisfies $Ric + \nabla^2 u \geq cI$ (equivalently, $[\nabla, l]f \geq c|\nabla f|^2$, where $[A, B] = AB - BA$). Then, (M, μ_u) has a log-Sobolev inequality with constant $\frac{1}{c}$.*

EXAMPLE 5.2. In finite-dimensional Gaussian space $(\mathbb{R}^n, \gamma_n)$, the Ricci curvature is zero and the Hessian of the potential $u = \frac{||x||^2}{2}$ is the identity: $\nabla^2 u = I$. Hence, we have a log-Sobolev inequality with $c = 1$.

EXAMPLE 5.3. In (S^n, σ^n), that is the sphere with Haar (uniform) measure, we have $u = 1$, so that $\nabla^2 u$ and the Ricci curvature is $Ric = \frac{n}{r^2} I$. Then, we have log-Sobolev inequality with $c = \frac{n}{r^2}$.

6. Parting Thoughts

One good thing about Poincaré and log-Sobolev inequalities is that they tensorise. That is, suppose we have the family $(X_i, \mu_i)_{i=1}^{\infty}$ with $X = \bigotimes_{i=1}^{\infty} X_i$ and $\mu = \otimes_i \mu_i$. Then, we have the following result:

THEOREM 6.0.1. *(Efron-Stein) Suppose our sequence of spaces above each has Poincaré inequalities with constants γ_i, then the space (X, μ) has Poincaré inequality with $\gamma = \inf_i \gamma_i$. Furthermore, if we define the entropy of f by $\mathrm{Ent}(f) = \int f \log \frac{f}{\int f d\mu} d\mu$, then we have the following entropy inequality:*

$$\mathrm{Ent}(f) \leq \sum_{i=1}^{\infty} \int \mathrm{Ent}_{\mu_i}(f) d\mu$$

CHAPTER 5

Lecture 5: PCA and Random Matrices
Scribe: K. Ramsey

1. Covariance Estimation and Random Matricies

We discuss covariance estimation, spiked matrix models and principle component analysis (PCA). Assume that $\{x_i\}_{i=1}^m$ are IID, centred $\mathbb{R}^n$ valued random vectors.

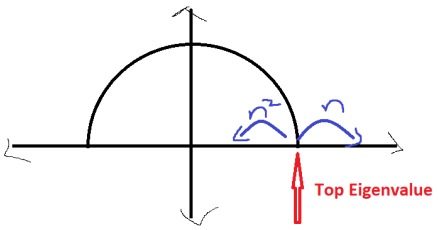
(1) Given the data, how can we estimate $\Sigma = \mathbb{E}XX^\top$?
(2) Given $\widehat{\Sigma}_m$, find the principle directions

2. Covariance Estimation

We know that $m^{-1}\widehat{\Sigma}_m$ is strongly consistent from the strong law of large numbers, but how fast is this happening? What is the sample complexity of this estimate? How large does $\alpha := m/n$ need to be in order for

$$||\widehat{\Sigma}_m - \Sigma||_{op} < \epsilon?$$

The answer is $\alpha \sim \epsilon^{-2}$, i.e., $m \sim n\epsilon^{-2}$, which is linear sample complexity. For an intuitive explanation of this rate, first recall that the operator norm of a matrix is the square root of the top eigenvalue. Now, remember that the empirical spectral distribution of a GOE approaches the semi-circle law. In order to shift the top eigenvalue to the right, we only need to move one eigenvalue. To move it to the left, we have to shift the entire spectrum. So, moving to the left should cost more: n^2 and moving to the right should cost less: n.



LEMMA 2.1. *Let A have IID centered, Gaussian entries,*

$$\Pr\left(||m^{-1}AA^\top - I||_{op} > C\left(\frac{1}{\sqrt{\alpha}} + \frac{t}{\sqrt{m}}\right)\right) \leq Ce^{-ct},$$

which implies that $\alpha \sim \epsilon^{-2}$, $m \sim n\epsilon^{-2}$.

PROOF. Technique: ϵ-net.
Choose an ϵ-net on S^{n-1} and call it $\mathcal{N}_\epsilon$. Remember that $|\mathcal{N}_\epsilon| \leq C(\epsilon)^n$. We show that the squared operator norm scales like n^2 :

$$\Pr\left(||m^{-1}AA^\top - I||_{op}^2 > t\right) \leq C(\epsilon)^n \sup_{v\in\mathcal{N}_\epsilon} \Pr\left(\left|\frac{||Av||^2}{m} - 1\right| > t\right).$$

Now, $m^{-1}||Av||^2$ is an IID sum of squared, standard Gaussian random variables; it is sub-exponential. Let $\delta = t/m + \lambda/\alpha$.

$$C(\epsilon)^n \sup_{v \in \mathcal{N}_\epsilon} \Pr\left(\left| \frac{||Av||^2}{m} - 1 \right| > t \right) \leq e^{n \log C(\epsilon) - Cm(\lambda/\alpha + \delta)} \leq e^{-Cm(\delta)}.$$

where λ is chosen so that the two terms cancel. □

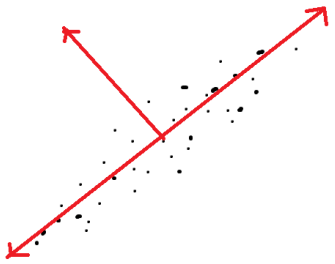
Can we do better than this? No. Suppose that $m/n \to \alpha$ as $m, n \to \infty$. If $||\widehat{\Sigma}_m - \Sigma||_{op} \to 0$ then the spectrum converges to the dirac mass at 1. We know that this does not happen from Marchenko-Pastur.

THEOREM 2.2 (Marchenko-Pastur). *Let $\mu_n = n^{-1} \sum \delta_{\lambda_i} \to \mu_\alpha$ weakly almost surely, where*

$$\mu_\alpha(dx) = \frac{1}{2\pi\alpha x} \sqrt{(\lambda_+ - x)(x - \lambda_-)} dx,$$
$$\lambda_\pm = \alpha(1 \pm \alpha^{-1/2})^2 \qquad \alpha \geq 1.$$

3. Spiked Matrix Models

Recall that in many applications there are a few important directions, or, principle directions and the rest is essentially noise. (These directions are the leading eigenvectors of Σ_n.) This phenomena suggests that we should model the data as noise plus some informative directions:



$$\Sigma_n = W_n + \sum_{\ell=1}^{k} \lambda_\ell v_\ell v_\ell^\top,$$

i.e., the data has a low dimensional linear substructure. We would like a model to reflect this; when we estimate the covariance we want this structure to appear. Consider the following models.

Spiked Wishart model: $x_i \sim N(0, I + \mu v v^\top)$. If $A = [x_1, \ldots, x_m]$ then $\widehat{\Sigma} = AA^\top + \lambda A v v^\top A^{-1}$. In the spiked Wishart model, for $\sqrt{\alpha}\lambda \leq 1$, we cannot really obtain any information about λ. When $\sqrt{\alpha}\lambda > 1$, something happens.

(1) $\sqrt{\alpha}\lambda \leq 1$: $|\langle v_{\max}, v_0 \rangle|^2 \to 0$
(2) $\sqrt{\alpha}\lambda > 1$: $|\langle v_{\max}, v_0 \rangle|^2 \to \frac{1 - \gamma/\beta^2}{1 - \gamma/\beta}$

Given $A_m = W_m + \lambda v v^\top$, we can estimate the covariance with

$$\frac{1}{m} \sum_{m=1}^{M} A_m = \frac{1}{m} \sum_{m=1}^{M} W_m + \lambda v v^\top.$$

(1) $\sqrt{\alpha}\lambda \leq 1$: $|\langle v_{\max}, v_1 \rangle|^2 \to 0$
(2) $\sqrt{\alpha}\lambda > 1$: $|\langle v_{\max}, v_1 \rangle|^2 \to 1 - \lambda^{-2}$

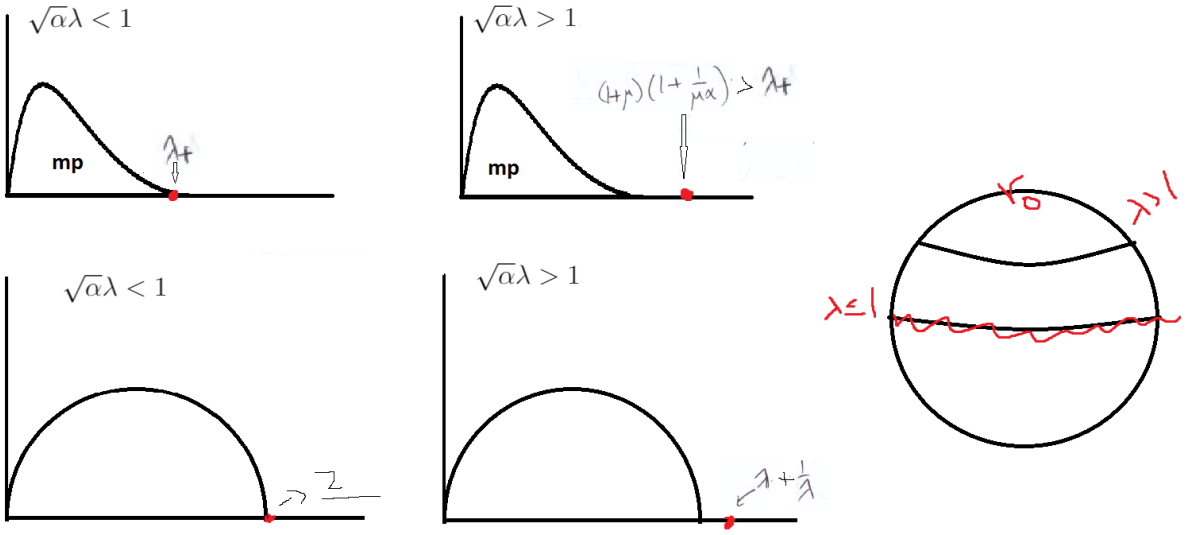


FIGURE 3.1. The BBP transition for spiked wishart and spiked GOE. When the SNR goes above a certain threshold, the top-eigenvalue separates from the bulk. Furthermore, the landscape of the corresponding function (Rayleigh quotient on the unit sphere) is such that below this threshold, all of the critical points (the eigenvectors) are near the equator, but above it, two of the critical points ($\pm$ the top EV) have macroscopic latitute.

CHAPTER 6

Lecture 6: Random Matrix Methods and the Semicircle Law
Scribe: K. Ramsay

1. The semi-circle and Marchenko–Pastur laws

Recall the BBP transition from last class. The goal of this lecture is to prove the semicircle law for Wigner ensembles and the MP for Gaussian with identity covariance.

THEOREM 1.1 (Wigner). *W_N is a $N \times N$ Wigner matrix: Hermitian matrix (symmetric in real valued case), with IID entries that are mean zero, $\mathbb{E}W_N(i,j)^2 = N^{-1}$ and finite $2+\epsilon$ moment, then*

$$\mu_N = \frac{1}{N}\sum_{\ell=1}^{N} \delta_{\lambda_i}. \to \mu_{sc} \text{ weakly, almost surely.}$$

THEOREM 1.2 (M.P). *If W_N is a $N \times N$ Wishart matrix: $W_N = AA^\top$ where A is matrix where each column is a standard N dimensional Gaussian vector and columns are independent, then,*

$$\mu_N = \frac{1}{N}\sum_{\ell=1}^{N} \delta_{\lambda_i}. \to \mu_{mp} \text{ weakly, almost surely.}$$

$$\mu_{mp}(dx) = \frac{1}{2\pi 1/\alpha x}\sqrt{(\lambda_+ - x)(x-\lambda_-)}dx,$$

$$\lambda_\pm = \alpha(1 \pm \alpha^{-1/2})^2 \qquad \alpha \geq 1.$$

One can take a combinatorial/ method of moments approach or a analytical approach which uses the Stieltjes transform. The Stieltjes transform method has a similar flavor to the Laplace/Fourier transform proof methods for the standard CLT.

2. Some Spectral Calculus

Recall that $A_{n\times n}$ Hermitian then $R_A(z) = (zI - A)^{-1}$ is the resolvent of A. This is a meromorphic, matrix valued function with poles at the spectrum of A. Recall that if $f\colon \mathbb{C} \to \mathbb{R}$ is analytic, then

$$f(A) = \frac{1}{2\pi i}\int_\Gamma \frac{f}{z-A}dz = \Sigma f(\lambda_i)v_i v_i^\top,$$

where Γ is a union of sets around the poles. This is the Cauchy integral transform. Now,

$$\frac{1}{N}\mathrm{tr}(f(A)) = \frac{1}{N}\mathrm{tr}(f(\lambda_i)) = \int f d\mu_N.$$

We can then access μ through the resolvent.

DEFINITION 2.1 (Stieltjes transform). Let μ be a positive measure on $\mathbb{R}$, then

$$S_\mu(z) = \int \frac{1}{z-x} d\mu(x), \qquad z \in \mathbb{C}\backslash\mathbb{R}$$

Observe that $S_\mu(z)$ is analytic on $\mathbb{C}\backslash\mathbb{R}$. Also, $S_\mu(z)$ characterizes μ.

THEOREM 2.2. *Let $I = (a,b) \subset \mathbb{R}$, with I being a continuity set of μ. Then,*

$$\mu(I) = \lim_{\epsilon \downarrow 0} \int_I \mathrm{Im}(S_\mu(z)) dz.$$

We note here the following useful facts.

LEMMA 2.3. *We have that:*

- *$\mu_N \to \mu$ weakly $\iff$ $S_{\mu_N(z)} \to S_\mu(z)$ pointwise on $\mathbb{C}_+$*
- *If μ is a probability measure, then if $S_{\mu_N(z)} \to S_\mu(z) \implies \mu_N \to \mu$ weakly.*
- *If μ_N are random and $S_{\mu_N(z)} \to S_\mu(z)$ for deterministic μ then $\mu_N \to \mu$ weakly in probability.*

3. Semicircle Law

THEOREM 3.1. *Let W_N be a GOE and $S_N(z) = S_{\mu_N}(z)$. Then, $S_N(z) \to S(z)$ which solves*

$$S(z) = \frac{1}{z - S(z)},$$

which is the unique S-transform of the semicircle.

These types of relations are sometimes called "self-consistent" or "mean field" equations.

PROOF. Step 1: Concentration! S_N is $\frac{1}{|\mathrm{Im}(z)|}^2$-Lipschitz. W_{ij} has variance $1/N$. Using Gaussian concentration we have that

$$\Pr(|S_N(z) - \mathbb{E}S_N(z)| > \epsilon) \leq Ce^{-cN\epsilon^2 \mathrm{Im}(z)^2}.$$

Step 2: Schur Complement Formula! Let W be $N \times N$ Hermitian. Let A^i be A with the i^{th} row and i^{th} column deleted. Let w_i be the i^{th} row of W with W_{ii} deleted. Then,

$$(z - W)^{-1}_{ii} = (z - W_{ii} - \langle w_i, R_{W^i}(z) w_i \rangle)^{-1}$$

and it follows that

$$S_N(z) = \frac{1}{N}\mathrm{tr}(R_W(z)) = \frac{1}{N}\sum_{i=1}^{N} (z - W_{ii} - \langle w_i, R_{W^i}(z) w_i \rangle)^{-1}.$$

Step 3: Concentration again with self consistency

$$S_N(z) \approx \mathbb{E}S_N(z) = \overline{S}_N(z).$$

Note that $w_i \perp T_{w^i}(z)$.

$$\begin{aligned}
\langle w_i, R_{W^i} w_i \rangle &\approx \mathbb{E}\langle w_i, R_{W^i} w_i \rangle \\
&\approx \mathbb{E}\mathrm{tr}(R_{W^i}) \frac{\mathbb{E}||W_i||^2}{||(1,\ldots,1)||^2} \\
&\approx \mathbb{E}S_{W^i} \approx \mathbb{E}S_W(z) = \overline{S}_N(z)
\end{aligned}$$

Then,

$$\overline{S}_N(z) \approx \frac{1}{N}\sum_{i=1}^{N}(z - W_{ii} - \overline{S}_N(z))^{-1}$$

$(W_{ii} = O(N^{-(1-\epsilon)}))$

$$\frac{1}{N}\sum_{i=1}^{N}(z - \overline{S}_N(z))^{-1} + O(N^{-(1-\epsilon)})$$

$$= (z - \overline{S}_N(z))^{-1} + o(1)$$

So, then

$$\overline{S} = \frac{\overline{S} - \sqrt{S^2 - 4}}{2}.$$

□

The steps of the above proof can be summarised by

(1) Concentration
(2) Schur Compliment
(3) Self consistency

Nect, we discuss the MP theorem.

4. Marchenko Pastur

For this proof, we have the 3 ingredients:

(1) Self consistency
(2) Concentration
(3) Schur Compliment -type

For 1. we have that $S_N \to S$ with

$$S(z - 1 + \alpha^{-1} - +\alpha^{-1}zS) = 1.$$

For 2. the issue is that the entries of W_N are not IID, but it has IID columns and so, we can use the following lemma.

LEMMA 4.1 (B-C-C). *If f has bounded variation with $\mathbb{E}\left|\int f d\mu_N\right| < \infty$, then*

$$\Pr\left(\left|\int f d\mu_N - \mathbb{E}\int f d\mu_N\right| > \delta\right) \leq 2e^{-N\delta^2\alpha/8}.$$

For 3. Use the Sherman-Morrison result: If A is invertible, then

$$(A + tvv^\top)^{-1}v = \frac{A^{-1}v}{1 + t\langle v, A^{-1}v\rangle}.$$

For the remaining details, see the lecture videos and notes.

CHAPTER 7

Lecture 7: The Marchenko–Pastur Law

1. Marchenko-Pastur

Our goal in this lecture is to prove the Marchenko-Pastur theorem.

THEOREM 1.1. *Let $(Y_\ell)_{\ell=1}^M$ be iid $N(0,I)$ Gaussians in $\mathbb{R}^N$ with $M=\alpha N$, and let*

$$W_N = \frac{1}{M}\sum Y_\ell Y_\ell^T$$

then for $\alpha > 1$ the empirical spectral measure $\mu_N = \frac{1}{N}\sum \delta_{\lambda_i}$ satisfies

$$\mu_N \to \mu_\alpha$$

weakly almost surely where

$$\mu_\alpha(dx) = \frac{\sqrt{(\lambda_+ - x)(x-\lambda_-)}}{2\pi x/\alpha}\mathbb{1}_{x\in[\lambda_-,\lambda_+]}dx$$

Recall from the proof of the semi-circle law that we had a three step strategy:

(1) Self-consistent equation
(2) Concentration
(3) Matrix identity

For us, the new self consistent equation will be the following.

LEMMA 1.2. [**2**]*For each $\alpha > 1$, the Stieltjes transform of the Marchenko -Pastur law, s_α is the unique solution to the self-consistent equation,*

$$s_\alpha(z)\left(z - 1 + \frac{1}{\alpha} - \frac{z}{\alpha}s_\alpha(z)\right) = 1 \tag{1.1}$$

Now for concentration, the issue is that the approach we had taken earlier, namely the view the stieltjes transform as a function of the entries and then use concentration of IIDs won't work because the matrix, W_N, doesnt have iid entires. To get around this we'll use the following fact.

LEMMA 1.3. *(Guntuboyina-Leeb, Bordenave-Caputo-Chafai* [**1**, Lemma C.1]*) If A is a random $N\times N$ hermitian matrix whose columns are independent then for any $f:\mathbb{R}\to\mathbb{R}$ with $\lim_{x\to\pm\infty} f(x) = 0$ and with $||f||_{TV} < \infty$ we have that for every $t \geq 0$, the empirical spectral measure μ_N satisfies*

$$P\left(|\int f d\mu_N - \mathbb{E}\int f d\mu_N| \geq t\right) \leq 2\exp\left(-\frac{Nt^2}{2||f||_{TV}}\right). \tag{1.2}$$

This will work for us since $x \mapsto (\gamma - x)^{-1}$ is a function which decays at infinity. Now for the matrix identity again things will have to change. The key point last time around was that the Schur complement formula let us break up the terms in the Stieljes transform in to individual terms which are independent, or of the form a random vector action on a random matrix both of whom are independent of each other. To get this in our case, we'll want to use the Sherman-Morisson formula: For any A invertible $N \times N$, $t \in \mathbb{R}$, and $v \in \mathbb{R}^N$ we have that

$$\left(A + tvv^T\right)^{-1} v = \frac{A^{-1}v}{1 + t\left\langle v, A^{-1}v\right\rangle} \tag{1.3}$$

We're now in the position to prove the main result.

Proof of Marchenko–Pastur Theorem. The key point in the following will be to show that the resolvent asymptotically solves the self-consistent equation (1.1). To this end, we'll do the following. As with the semi-circle law, observe that for any z it suffices to show the result for the mean by (1.2). For ease of notation, lets denote $\bar{s}_N(z) = \mathbb{E}s_N(z)$. Our starting point is to observe that

$$R_{W_N}(z)W_N = \frac{1}{M}\sum_{\ell=1}^{M} R_{W_N}(z)Y_\ell Y_\ell^T.$$

Now, observe that we may wrote

$$R_{W_N}(z) = (z - W_N)^{-1} = \left(z - W_N^\ell - \frac{1}{M}Y_\ell Y_\ell^T\right)^{-1}$$

where $W_N^\ell = W_N - \frac{1}{M}Y_\ell Y_\ell^T$. Thus applying (1.3) with $A = z - W_N^\ell$, $t = -\frac{1}{M}$ and $v = Y_\ell$ we get that

$$R_{W_N}(z)Y_\ell = \frac{R_{W_N^\ell}(z)Y_\ell}{1 - \frac{1}{M}\left\langle Y_\ell, R_{W_N^\ell}(z)Y_\ell\right\rangle}.$$

Now, using the concentration argument from above we have

$$\begin{aligned}
\frac{1}{M}\left\langle Y_\ell, R_{W_N^\ell}(z)Y_\ell\right\rangle &\approx \frac{1}{M}\mathbb{E}\left\langle Y_\ell, R_{W_N^\ell}Y_\ell\right\rangle \\
&\approx \frac{1}{M}\mathbb{E}\operatorname{tr} R_{W_N^\ell} \\
&\approx \frac{1}{M}\mathbb{E}\operatorname{tr} R_{W_N}(z) = \frac{1}{\alpha}\bar{s}_N(z)
\end{aligned}$$

where in the second to last line we're using the rotation invariance of W_N^ℓ and the fact that $||Y_\ell||$ concentrates about $\sqrt{N}$ and in the last line we're using that W_N^ℓ is essentially W_N. Now if we combine these results we have that

$$\frac{1}{M}\sum_\ell R_{W_N}Y_\ell Y_\ell^T \approx \frac{1}{1 - \frac{1}{\alpha}\bar{s}(z)}\frac{1}{M}\sum_\ell R_{W_N^\ell}Y_\ell Y_\ell^T$$

So if we take expectations, we get that

$$\begin{aligned}\mathbb{E}R_{W_N}W_N &\approx \frac{1}{1-\frac{1}{\alpha}\bar{s}(z)}\frac{1}{M}\sum_{\ell}\mathbb{E}R_{W_N^\ell}Y_\ell Y_\ell^T \\ &= \frac{1}{1-\frac{1}{\alpha}\bar{s}(z)}\frac{1}{M}\sum_{\ell}\mathbb{E}R_{W_N^\ell}\cdot\mathbb{E}Y_\ell Y_\ell^T \\ &= \frac{1}{1-\frac{1}{\alpha}\bar{s}(z)}\mathbb{E}R_{W_N^1}\end{aligned}$$

on the otherhand, we have that $-1+zR_{W_N} = R_{W_N}W_N$ so that if we take an expectation of this inequality and $\frac{1}{N}\operatorname{tr}$, we ge

$$-1+z\bar{s}_N(z) \approx \frac{1}{1-\frac{1}{\alpha}\bar{s}_N(z)}\bar{s}_N(z).$$

Re-arranging this yields

$$\bar{s}_N(z)\left(z-1+\frac{1}{\alpha}-\frac{z}{\alpha}\bar{s}_N(z)\right)\approx 1$$

which will give us what we need as $N\to\infty$ □

CHAPTER 8

Lec 8: A Short Proof of the BBP transition for GOE and Gaussian Wishart ensembles

1. Introduction

In this note we consider the following two random matrix ensembles $W_N \sim GOE$ which is the real symmetric Gaussian ensemble and W_N which is a Wishart distribution of the form $W_N = V_N V_N^T$ where $V_N = [g_1, \ldots, g_M]$, and $M/N \to \alpha$. We give here a short proof via deterministic identities assuming convergence of the empirical spectral distributions.

2. A matrix identity

Let A be a symmetric (Hermitian) $N \times N$ matrix with eigenvalues by $\lambda_1 \geq \ldots \lambda_N$ and corresponding eigenvectors $v_1, \ldots, v_N$. Let $Q_A(\gamma, w) = \langle w, (\gamma - A)^{-1} w \rangle$, denote the Rayleigh quotient of the resolvent of A. Observe that if we restrict the domain of $Q_N(\cdot, w)$ to the $\gamma \in (\lambda_+, \infty)$ where $\lambda_+ = \max \{\lambda_i : \langle v_i, w \rangle \neq 0\}$, then $Q_A(\cdot, w)$ is smooth and monotone decreasing with

$$\lim_{\gamma \downarrow \lambda_+} Q_A(\gamma, w) = \infty$$
$$\lim_{\gamma \uparrow \infty} Q_A(\gamma, w) = 0.$$

In particular Q_A is invertible in $(\lambda_+(w), \infty)$,with inverse $Q_A^{-1}(\cdot, w)$. Consider now a rank one perturbation of A: let $\mu > 0$, and $v \in \mathbb{R}^N$, and let

$$L_A(\mu) = \lambda_1(A + \mu v v^T),$$

denote the top eigenvalue of $A + \mu v v^T$. We then have the following explicit characterization of $L(\mu)$

LEMMA 2.1. *Let A be a symmetric (Hermitian) $N \times N$ matrix, $\mu > 0$, and $v \in \mathbb{R}^N$. Then $A + \lambda v v^T$ has $L_A(\mu) = Q_A^{-1}(\frac{1}{\mu}, v) \vee \lambda_1(A)$. Furthermore L is non-decreasing, convex and differentiable except at at most one point.*

PROOF. For clarity of notation, we suppress subscripts of A. Let us diagonalize A and write

$$Q(\gamma, v) = \sum_i \frac{v_i^2}{\gamma - \lambda_i}.$$

Observe then that the characteristic polynomial $p(\lambda)$ of $A + \mu v v^T$has

$$\begin{aligned} p(\lambda) &= \det(\lambda I - A - \mu v v^T) \\ &= \det(\lambda I - A) - \mu \langle v, \mathrm{adj}(\lambda - A) v \rangle . \\ &= \det(\lambda I - A) (1 - \mu Q(\gamma, v)) \end{aligned}$$

where $\text{adj}(B) = \det(B) \cdot B^{-1}$. By definition of Q and Q^{-1} we see that the characteristic polynomial of $A+\lambda vv^T$ is that of A times an additional factor which is an algebraic function whose poles are at those of A. Evidently this yields

$$L(\mu) = Q^{-1}(1/\mu) \vee \max_{i \in I^c} \lambda_i = Q^{-1}(1/\mu) \vee \max_i \lambda_i$$

which gives us the identity.

The functional properties are self-evident: write L as a Rayleigh quotient we see that

$$L(\mu) = \max_{||x||=1} \left\langle x, \left(A + \mu vv^T\right) x \right\rangle,$$

is a maximum of linear functions so that it is convex. The differentiability is clear from that of Q^{-1} so that it smooth except for one point which much satisfy $Q^{-1}(\frac{1}{\mu_0}) = \lambda_1(A)$. □

It will be helpful to also note the following corollary. Let $v_1(\mu)$ denote a top eigenvector of $A + \mu vv^T$

COROLLARY 2.2. *Suppose that $v \cdot v_1 \neq 0$ then*

$$\langle v_1(\mu), v\rangle^2 = L'(\mu).$$

Suppose that $v_1 = 0$ and let μ_0 be the unique point such that $Q(\lambda_1, v) = 1/\mu_0$. If $\mu \neq \mu_0$ then the preceding holds and at μ_0 we have

$$\partial_- L'(\mu_0) \leq \langle v_1(\mu), v\rangle^2 \leq \partial_+ L'(\mu_0),$$

PROOF. The first result follows by the envelope theorem: since

$$L(\mu) = \max_{||x||_2=1} \left\langle x, (A + \mu vv^T)x \right\rangle,$$

we have $L'(\mu) = \langle v_1(\mu), v\rangle^2$. The rest follows by continuity. □

COROLLARY 2.3. *The above holds provided $A + \lambda uv^T$ is similar to a symmetric matrix, $B = \Sigma^{-1}(A + \lambda uv^T)\Sigma$ where $Q(\lambda, u, v) = \langle u, R_A(\lambda)v\rangle$, except that if we let w_1 be the top EV for B thenwe ge*

$$L'(\mu) = \left\langle w_1, \Sigma^{-1}u \right\rangle \langle v, \Sigma w_1\rangle.$$

PROOF. For clarity of notation, we suppress subscripts of A. Since $A + \lambda uv^T$ is similar to a symmetric matrix, i.e., $B = \Sigma^{-1}\left(A + \lambda uv^T\right)\Sigma$ is symmetric for some invertible Σ. In particular $P = A + \lambda uv^T$ has n eigenvalues and n eigenvectors $v_i = \Sigma w_i$ where w_i are the eigenvectors of B so we may write

$$Q(\gamma, u, v) = \sum_i \frac{\langle u, v_i\rangle \langle v, v_i\rangle}{\gamma - \lambda_i}.$$

Observe then that the characteristic polynomial $p(\lambda)$ of $A + \mu vv^T$has, as before,

$$\begin{aligned} p(\lambda) &= \det(\lambda I - A - \mu uv^T) \\ &= \det(\lambda I - A)\left(1 - \mu Q(\gamma, u, v)\right). \end{aligned}$$

this yields the relevant identity. To the derivative of $L(\mu)$ note that since we have the similarity propery, we may write

$$L(\mu) = \max \langle x, Bx\rangle$$

so differentiating we get

$$\begin{aligned} L'(\mu) &= \langle w_1, \partial_\mu B w_1 \rangle \\ &= \left\langle w_1, \partial_\mu \left(\Sigma^{-1}(A + \mu u v^T)\Sigma \right) w_1 \right\rangle \\ &= \left\langle w_1, \Sigma^{-1} u \right\rangle \langle v, v_1 \rangle \end{aligned}$$

□

3. BBP Transition

THEOREM 3.1. *(Wigner BBP) Consider*

$$A_N = W_N + \beta v v^T$$

hen we ahve that the top EV is given, asympotically, by

$$\lambda_1(\beta) = \begin{cases} 2 \\ \frac{1}{\beta} + \beta \end{cases}$$

PROOF. By rotational invariance of the law of W_N, we may take $v = (1, \dots, 1)/\sqrt{N}$ without loss of generality Then

$$Q_N(z, v) = \frac{1}{N} Tr(z - W_N)^{-1} = s_N(z)$$

is the Stieltjes transform of the empirical spectral distribution of W_N. Note then that

$$s_N(z) \to s(z)$$

which solves

$$s + \frac{1}{s} = z.$$

We also have by the Kolmos-Furedi theorem that

$$\lambda_1(W_N) \to 2$$

almost surely. On the otherhand by the preceding we see that

$$\lambda_1(A_N) = \gamma_N(\beta) \vee \lambda_1(W_N)$$

where $\gamma_N(\beta)$ solves $Q(\gamma_N(\beta), v_1^N) = \frac{1}{\beta}$, i.e., it solves

$$s_N(\gamma_N(\beta)) = \frac{1}{\beta}$$

but since $s_N(z) + \frac{1}{s_N}(z) = z + o(1)$, we must have that

$$\frac{1}{\beta} + \beta = \gamma_N(\beta) + o(1)$$

sending $N \to \infty$ gives $\gamma_N(\beta) \to \frac{1}{\beta} + \beta$. Similarly $L_N(\mu) \to L(\mu)$ which can be shown to be convex and differentiable for $\beta \neq 1$ so that

$$L_N(\gamma_N(\beta)) \to \frac{1}{\beta} + \beta$$

and by Griffth's lemma for the convergence of derivatives of convex functions,

$$L_N'(\gamma_N(\beta)) \to L'(\beta) = 1 - \frac{1}{\beta^2}$$

as desired. □

THEOREM 3.2. *(BBP Wishart) We have that if we let* $\Sigma_N(\lambda) = I + \beta vv^T$ *then*

$$P_N = \frac{1}{M}\sqrt{\Sigma_N(\beta)}A_N A_N^T\sqrt{\Sigma_N(\beta)}$$

has top eval going to

$$\lambda_1(\beta) = \begin{cases} (1+\frac{1}{\sqrt{\alpha}})^2 & \sqrt{\alpha}\beta < 1 \\ (\beta+1)\left(1+\frac{1}{\beta\alpha}\right) & \sqrt{\alpha}\beta > 1 \end{cases}$$

CHAPTER 9

Lec 9: Comparison Inequalities
Scribe: M. Majid

1. Introduction

The goal of today's class is to try to understand, and give a fast introduction to comparison inequalities. Comparison inequalities are very powerful tools for comparing properties of one class of stochastic processes to another. Largely—if not entirely— the theory is built around Gaussians.

Suppose two centered Gaussian processes are given, $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$, where T is a finite set. For simplicity let's assume these are Gaussian random vectors and T is $[n]$. The question Gaussian comparison theory wants to ask is the following.

Question. Suppose some smooth function $F : \mathbb{R}^n \to \mathbb{R}$ is given, and you can compute the expected value of this function with respect to these distributions, that is $\mathbb{E}[F(X)]$ and $\mathbb{E}[F(Y)]$. The question of Gaussian comparison theory is whether we can show that one of these expected values is greater than the other,

$$\mathbb{E}[F(X)] \leq \mathbb{E}[F(Y)].$$

In this lecture, we largely focus on the question where F is the maximum, that is,

$$F(X) = \sup_{t\in T} X_t.$$

There is a much more general theory that holds for more complicated functions. Informally, where the theory breaks in generalizing from Gaussians to non-Gaussians, is in this extension to more general functions.

Before getting into the heart of the matter let's review some useful facts about Gaussians.

2. Some Useful Facts About Gaussians

2.1. Integration by Parts for Gaussians.

LEMMA 2.1. *Suppose $z \sim \mathcal{N}(0, \Sigma)$, where $\Sigma \succ 0$, and $z \in \mathbb{R}^n$ Let $v \in \mathbb{R}^n$, and F be a smooth function with polynomial growth at infinity. Then*

$$\mathbb{E}[D_v F(z)] = \mathbb{E}[F(z)\langle \Sigma^{-1} z, v\rangle].$$

In other words, the expected value of the directional derivative of F in the direction of v, evaluated at z, is equal to the expected value of F at z times the inner product of the inner product of z and v, with respect to the metric induced by the inverse covariance matrix.

The idea of the proof is, if we take a Gaussian random vector and we add another vector to it, then the new Gaussian is a Gaussian with a new mean, but as long as you preserve the covariance structure, changing the mean amounts to the tilt of the measure.

PROOF.

$$\mathbb{E}[D_v F(z)] = \lim_{h\to 0} \mathbb{E}\left[\frac{F(z+hv) - F(z)}{h}\right]$$

Now by Cameron-Martin-Girsanov theorem we know that

$$\mathbb{E}[F(z+hv)] = \mathbb{E}\left[F(z)e^{h\langle\Sigma^{-1}z,v\rangle - \frac{h^2}{2}\langle\Sigma^{-1}v,v\rangle}\right]$$

Therefore we just need to plug this theorem in the above equation.

$$\mathbb{E}[D_v F(z)] = \lim_{h\to 0} \mathbb{E}\left[\frac{F(z+hv) - F(z)}{h}\right] = \lim_{h\to 0} \mathbb{E}\left[F(z)\frac{e^{h\langle\Sigma^{-1}z,v\rangle + O(h^2)} - 1}{h}\right] = \mathbb{E}[F(z)\langle\Sigma^{-1}z, v\rangle]$$

□

COROLLARY 2.2. *Under the same setting, but more generally for* $\Sigma \succcurlyeq 0$*,*

$$\mathbb{E}[z_1 F(z)] = \sum_j \mathbb{E}[z_1 z_j] \cdot \mathbb{E}[\partial_j F(z)]$$

PROOF. If $\Sigma \succ 0$ apply Lemma 2.1 with $v = \Sigma e_1$ □

2.2. Interpolation Inequality and Slepian's Lemma.

THEOREM 2.2.1 (Slepian). *Suppose* $(X_t) \sim \mathcal{N}(0, \Sigma_X), (Y_t) \sim \mathcal{N}(0, \Sigma_Y)$*, and*

$$\begin{aligned}
\Sigma_X(s,t) &\le \Sigma_Y(s,t) \quad (s,t) \in A \\
\Sigma_X(s,t) &\ge \Sigma_Y(s,t) \quad (s,t) \in B \\
\Sigma_X(s,t) &= \Sigma_Y(s,t) \quad (s,t) \in (A \cup B)^c.
\end{aligned}$$

If we have some smooth function $F \in C^\infty$ *with*

$$\begin{aligned}
\partial^2_{st} F &\ge 0 \quad (s,t) \in A \\
\partial^2_{st} F &\le 0 \quad (s,t) \in B \\
\partial^2_{st} F &= 0 \quad (s,t) \in (A \cup B)^c,
\end{aligned}$$

then

$$\mathbb{E}[F(X)] \le \mathbb{E}[F(Y)].$$

In other words, if the signs of the entries of the Hessian of our function always match up the sign of the difference between the two covariances then we have the comparison inequality we are shooting for.

PROOF. The idea is to use the smart path method again. Define

$$Y_\tau = \sqrt{1-\tau}X + \sqrt{\tau}Y,$$

which is a processes that at time $\tau = 0$ is equal to X and at time $\tau = 1$ is equal to Y. Now let's look at the function

$$\phi(\tau) = \mathbb{E}[F(Y_\tau)].$$

Our goal is to show $\phi'(t) \ge 0$, because if this holds then we'll have

$$\mathbb{E}[F(Y)] = \phi(1) \ge \phi(0) = \mathbb{E}[F(X)],$$

as desired. By taking the derivative we get

$$\begin{aligned}\phi'(t) &= \partial_\tau \mathbb{E}[F(Y_\tau)] \\ &= \mathbb{E}[\langle \nabla F(Y_\tau), \dot{Y}_\tau \rangle] \\ &= \sum_s \mathbb{E}[\partial_s F(Y_\tau) \dot{Y}_\tau^{(s)}]\end{aligned}$$

Now we can think of the Gaussian process this way: take the vector Y_τ and augment it at the beginning by $\dot{Y}_\tau^s$. Then

$$\begin{aligned}\phi'(t) &= \sum_s \mathbb{E}[\partial_s F(Y_\tau) \dot{Y}_\tau^s] \\ &= \sum_s \left[\sum_t \mathbb{E}\left[\dot{Y}_\tau^s Y^t\right] \cdot \mathbb{E}\left[\partial_{ts}^2 F(Y_\tau)\right] \right].\end{aligned}$$

Now note that the covariance term

$$\mathbb{E}[\dot{Y}_\tau^s Y_\tau^t] = \frac{1}{2}(\Sigma_Y(t,s) - \Sigma_X(t,s)),$$

Therefore

$$\begin{aligned}\phi'(t) &= \sum_s \left[\sum_t \mathbb{E}\left[\dot{Y}_\tau^s Y^t\right] \cdot \mathbb{E}\left[\partial_{ts}^2 F(Y_\tau)\right] \right] \\ &= \left[\sum_{(s,t)\in A} (+)(+) \right] + \left[\sum_{(s,t)\in B} (-)(-) \right] + 0 \geq 0,\end{aligned}$$

as desired. □

Remark 2.3. As you can see in the last step of the proof we only require one of the zero conditions,

$$\partial_{st}^2 F = 0 \quad (s,t) \in (A \cup B)^c,$$

or

$$\Sigma_X(s,t) = \Sigma_Y(s,t) \quad (s,t) \in (A \cup B)^c.$$

Now we want to focus on understanding a bit better, the suprema of stochastic processes, specifically, the suprema of Gaussian processes.

Theorem 2.3.1 (Slepian's Lemma). *Suppose the two Gaussian processes (X_t), (Y_t) have the same variances, $\mathbb{E}[X_t^2] = \mathbb{E}[Y_t^2]$, and if we look at the second moment of the increments we have*

$$\mathbb{E}[(X_t - X_s)^2] \leq \mathbb{E}[(Y_t - Y_s)^2].$$

Then we have

$$\forall s > 0, \quad \Pr[\max_t X_t > s] \leq \Pr[\max_t Y_t > s]. \tag{2.1}$$

Furthermore,

$$\mathbb{E}[\max_t X_t] \leq \mathbb{E}[\max_t Y_t]. \tag{2.2}$$

Outline of the proof. We want to use the interpolation inequality 2.2.1, which is of the flavor of $\mathbb{E}[f(x)] \leq \mathbb{E}[f(y)]$. The first thing is to unpack the assumptions of this Theorem

$$\mathbb{E}[X_t^2] = \mathbb{E}[Y_t]^2, \mathbb{E}[X_t^2 - X_s^2] \leq \mathbb{E}[Y_t^2 - Y_s^2] \implies \mathbb{E}[Y_s Y_t] \leq \mathbb{E}[X_t X_s],$$

which is an inequality comparing the covariances of the two stochastic processes. Now the trick is to take a function $f(x) \approx \mathbb{1}[\max_i x_i < s]$. Note that if the maximum is smaller than s, then all of the entries are. Then our goal is to prove $\mathbb{E}[f(X)] \geq \mathbb{E}[f(Y)]$, since after subtracting a 1 and negation, this gives us the desired inequality, $\Pr[\max_t X_t > s] \leq \Pr[\max_t Y_t > s]$.

The idea is that if we let h be a smooth approximation of the indicator function,

$$h(x) \approx \mathbb{1}_{(-\infty, s)}(x),$$

then

$$f(x) = \prod_i h(x_i),$$

since as we mentioned before

$$\max_i x_i < s \iff \forall i : x_i < s.$$

Now we should look at the second derivative of f which is

$$\partial_{ij} f = h'(x_i) h'(x_j) \prod_{\ell \neq i,j} h(x_\ell).$$

Now if h is kind of nice for example something like Figure 2.1, then $h'(x_i)$ is always non-positive and $h(x_i)$ is always non-negative. Therefore

$$\partial_{ij} f = \overbrace{h'(x_i)}^{\leq 0} \overbrace{h'(x_j)}^{\leq 0} \prod_{\ell \neq i,j} \overbrace{h(x_\ell)}^{\geq 0} \geq 0,$$

and therefore by applying Theorem 2.2.1 we obtain $\mathbb{E}[f(X)] \geq \mathbb{E}[f(Y)]$ which gives us the desired bound. □

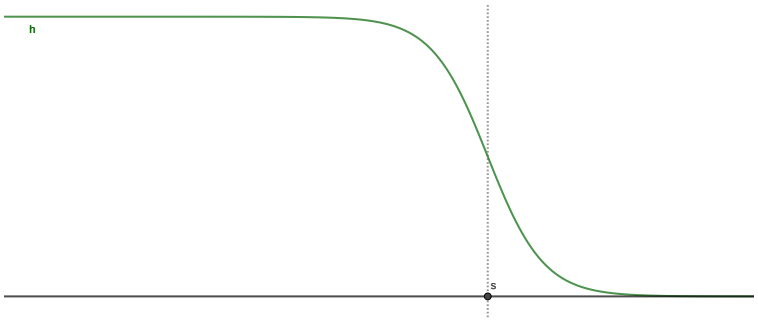


FIGURE 2.1. h is a smooth indicator function.

2.3. Sudakov Fernique inequality. The issue with Theorem 2.3.1 is that the constraints are still annoying to check, especially the equality of variances constraint $\mathbb{E}[X_t^2] = \mathbb{E}[Y_t^2]$. Unfortunately we cannot give up this constraint and get the probability bound 2.1, but we can relax this condition if we are willing just to get the inequality about the expectations 2.2. This motivates us to study the Sudakov-Fernique ineqaulity.

THEOREM 2.3.2 (Sudakov-Fernique). *Suppose we have two Gaussian processes $(X_t), (Y_t)$ and the second moment increment of X is dominated by the second moment increment of Y, i.e.*

$$\mathbb{E}[(X_t - X_s)^2] \leq \mathbb{E}[(Y_t - Y_s)^2].$$

Then

$$\mathbb{E}[\max_t X_t] \leq \mathbb{E}[\max_t Y_t]$$

Outline of the proof. The idea of the proof is to use a clever approximation of the maximum function. This idea is motivated form Statistical Physics and in the Physics community it goes by the name *"annealing"*.

The idea of annealing is the following statement. If we take the maximum of n things, i.e. $\max_i x_i$, by itself, this function is not smooth but it is $1-$Lipschitz. However, if we take

(Free Energy) $$\max_i x_i \leftarrow \frac{1}{\beta} \log \sum_i e^{\beta x_i},$$

also known as the free energy, then it is not hard to prove that

$$\max_i x_i \leq \frac{1}{\beta} \log \sum_i e^{\beta x_i} \leq \max_i x_i + \frac{\log n}{\beta}.$$

Therefore, we obtained a smooth approximation of the maximum function, as desired.

One proof of the Sudakov-Fernique inequality will be to use to use free energy as our function and some work will give us the desired inequality. □

Now let's see a few applications of this inequality.

EXAMPLE 2.4. Let A be an $m \times n$ Gaussian matrix (each entry of $A \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$). Then

$$||A||_{2,2} = \sup_{||u||=1, ||v||=1} \langle v, Au \rangle \implies \mathbb{E}||A||_{2,2} \leq \sqrt{m} + \sqrt{n}.$$

This is a good observation because if we take $n = m$ and let G be the GOE matrix

$$\mathbb{E}||G||_{op} = \mathbb{E}\frac{||A||_{2,2}}{\sqrt{n}} \leq 2,$$

which great because 2 is what we would predict from the semi-circle law. Therefore, this inequality together with the Borell's inequality already give us a proof of the Kalmos-Furedi result that we mentioned in the last lecture.

PROOF. Suppose $T = S^{n-1} \times S^{m-1}$, then

$$\mathbb{E}[\sup_{u \in S^{n-1}, v \in S^{m-1}} \langle v, Au \rangle] = \mathbb{E}[\sup_{(u,v) \in T} g(u,v)],$$

where $g(u,v) = \langle v, Au\rangle$. Note that g is a Gaussian process. Now if we look at the second moment increment of g we have

$$\begin{aligned}\mathbb{E}\left[(g(u,v)-g(x,y))^2\right] &= \mathbb{E}\Big[\sum_{i,j} A_{i,j}(v_iu_j - y_ix_j)\Big]^2 \\ &= \sum_{i,j}(v_iu_j - y_ix_j)^2 \\ &= ||u\otimes v - x\otimes y||_F^2 \\ \text{(Triangle Inequality)} \qquad &\leq ||u-x||_2^2 + ||v-y||_2^2.\end{aligned}$$

Now we argue that the last term is just the sum of the second moment increments of two *Isonormal processes* (Definition 2.5). To see this, let's look at the covariance between u and x with respect to the Isonormal process

$$\begin{aligned}\mathbb{E}[h(u)h(x)] &= \mathbb{E}[\langle g', u\rangle \cdot \langle g', x\rangle] \\ &= \langle u, x\rangle,\end{aligned}$$

and therefore

$$\mathbb{E}\left[(h(u)-h(x))^2\right] = ||u-x||_2^2,$$

as desired. Now let's get back to the problem.

$$\mathbb{E}\left[(g(u,v)-g(x,y))^2\right] \leq \underbrace{||u-x||_2^2}_{\text{increment of Isonormal in } \mathbb{R}^n} + \underbrace{||v-y||_2^2}_{\text{increment of Isonormal in } \mathbb{R}^m}$$

Now let $\tilde{g}(u,v) = h(u) + h'(v)$, where h, h' are independent Isonormal processes in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Using this process we get

$$\begin{aligned}\mathbb{E}\left[(g(u,v)-g(x,y))^2\right] &\leq ||u-x||_2^2 + ||v-y||_2^2 \\ &= \mathbb{E}\left[(\tilde{g}(u,v)-\tilde{g}(x,y))^2\right].\end{aligned}$$

Now if we apply Sudakov-Fernique 2.3.2 for g and $\tilde{g}$ we obtain

$$\mathbb{E}||A||_{2,2} \leq \mathbb{E}[\max_T \tilde{g}] \leq \mathbb{E}[\max_{u\in S^{n-1}} h(u)] + \mathbb{E}[\max_{v\in S^{m-1}} h'(v)]. \tag{2.3}$$

Now we should to bound each term separately.

$$\mathbb{E}[\max_{u\in S^{n-1}} h(u)] = \mathbb{E}[\max_{||u||=1}\langle h, u\rangle] = \mathbb{E}||h||_2 = \sqrt{n}.$$

Similarly for h', $\mathbb{E}[\max_{v\in S^{m-1}} h'(v)] = \sqrt{m}$. Therefore by 2.3 we may obtain

$$\mathbb{E}||A||_{2,2} \leq \sqrt{n} + \sqrt{m},$$

as desired. □

Definition 2.5 (Isonormal process, Canonical process). An Isonormal process on $\mathbb{R}^d$ is a Gaussian process that takes in a vector on outputs of the inner product of that vector with the standard Gaussian in $\mathbb{R}^d$, i.e.

$$\forall v \in \mathbb{R}^d : h(v) = \langle g', d\rangle, \quad g' \sim \mathcal{N}(0, I_d)$$

CHAPTER 10

Lec 10: Recovering a Vector From a Few Random Measurements
Scribe: M. Majid

1. Introduction

In this lecture we aim to study the problem of recovering a signal from a few (linear) measurements. What do we mean by this? Consider the following problem.

2. Problem

Suppose there exists some unknown parameter $x_0 \in \mathbb{R}^d$ which we don't have access to. Instead, we have access to $(y_\ell, a_\ell)_{\ell=1}^n$ where $y_\ell \in \mathbb{R}, a_\ell \in \mathbb{R}^d$ and we know a priori that $y\ell = a_\ell \cdot x_0$. Our goal is to obtain an estimate for x_0.

Let $A = [a_1 \cdots a_n]^T$. If $n = d$ and A is invertible this problem is easy. The setting we are interested in today is the case where $n < d$. For example one case can be the case where n is a constant. A more general case is the case where n and d are diverging in a way that $\frac{n}{d} < 1$.

3. Assumptions

The first move here is to put some a priori information about x_0. Otherwise, one may just solve the under-determined equation $y = Ax$, which has a subspace of solutions and one may obtain arbitrarily large solution by adding vectors from $\ker(A)$ to a solution. Therefore, we need to put some a priori information about x_0. One may think of this as a *penalty*.

We may impose different assumptions. One that is particularly interesting to us is the assumption that x_0 is in some compact convex set $\mathcal{K} \subset \mathbb{R}^d$. If we add this assumption, then one way of solving this problem is

$$(*) \qquad \begin{aligned} \text{find} \quad & x, \\ \text{such that} \quad & y = Ax, \\ & x \in \mathcal{K}. \end{aligned}$$

We can produce a solution to this problem efficiently through convex optimization methods [1] One method is to define $\ell(x) = ||y - Ax||_2^2$ and minimize it over $\mathcal{K}$. The issue with this approach is that in this method we will not necessarily get a point that satisfies $y = Ax$. Another method has to deal with the fact that compact convex sets are in correspondence with norms.

[1] Not directly to this problem, but an equivalent of it that we describe in Section 4.

4. Gauge Function and The Equivalent of $*$

DEFINITION 4.1 (Gauge Function). The gauge function of a set $\mathcal{K}$ takes in a vector x and outputs the smallest constant r, such that x/r is still in $\mathcal{K}$. Formally,

$$\mathcal{P}_{\mathcal{K}}(x) = \inf \{r > 0 : \frac{x}{r} \in \mathcal{K}\}.$$

Observations. We have two important observations here.

(1) $\mathcal{P}_{\mathcal{K}}(x) \leq 1 \iff x \in \mathcal{K}$.
(2) $\mathcal{P}_{\mathcal{K}}(x)$ is a (semi)-norm:
   (a) $\mathcal{P}_{\mathcal{K}}(x+y) \leq \mathcal{P}_{\mathcal{K}}(x) + \mathcal{P}_{\mathcal{K}}(y)$.
   (b) $\mathcal{P}_{\mathcal{K}}(ax) = |a|\mathcal{P}_{\mathcal{K}}(x)$.
   (c) $\mathcal{P}_{\mathcal{K}}(x) \geq 0$.

The issue with $\mathcal{P}_{\mathcal{K}}$ not being a norm is that in norms we have $\mathcal{P}_{\mathcal{K}}(x) = 0 \iff x = 0$, which is not the case for general $\mathcal{K}$s. In order for $\mathcal{P}_{\mathcal{K}}$ to be a norm, we should impose some *nice* assumptions for $\mathcal{K}$, for example $\mathcal{K}$ containing a neighborhood around the origin should be good enough.

Notation. We denote the (semi)-norm corresponding to $\mathcal{P}_{\mathcal{K}}$ by $||x||_{\mathcal{K}} = \mathcal{P}_{\mathcal{K}}(x)$.

Having defined the norm corresponding to the $\mathcal{K}$, we can finally say how we want to solve this problem. The approach is to solve this optimization problem:

$$\text{(equivalent of } *) \qquad \begin{aligned} \text{minimize} \quad & ||x||_{\mathcal{K}}, \\ \text{such that} \quad & y = Ax. \end{aligned}$$

Unlike minimizing $\ell(x) = ||y - Ax||_2^2$, we now know that $y = Ax$. Furthermore, we know that the solution is going to be inside the set $\mathcal{K}$, as $x_0 \in \mathcal{K}$, and therefore the minimum is at most 1. Note that we can solve this optimization problem efficiently using convex optimization methods.

What can we tell about the solutions to the problem above? Are they actually decent approximations of x_0? If $\mathcal{K}$ is not compact, then our solution may be arbitrarily far from x_0 while persevering the equality. Now that $\mathcal{K}$ is compact and convex, our concern is that it might be the case that x and x_0 are both inside $\mathcal{K}$ but are far from each other (see Figure 4.1).

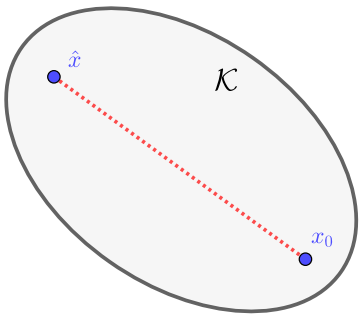


FIGURE 4.1. Even with the guarantee of x_0 being inside the convex compact set $\mathcal{K}$, our solution to $*$ might be far from x_0.

But if we choose our measurements are done randomly, then we can prove that our solution is close to x_0. In the following sections we aim to establish the tools for proving

Theorem 5.2.1 and finally prove it. This Theorem states a bound for the distance between the solution of Problem $*$ and the unknown parameter where the measurements are done randomly.

5. Gaussian Width and Spherical Width

DEFINITION 5.1 (Gaussian Width). The Gaussian Width of $T \subseteq \mathbb{R}^d$ is

$$w(T) := \mathbb{E}[\sup_{v \in T} \langle g, v \rangle], \quad g \sim \mathcal{N}(0, I_d).$$

REMARK 5.2. To tie this to the last lecture, note that $g(v) = \langle g, v \rangle$ is just the canonical Gaussian (isonormal) process on $\mathbb{R}^d$. Therefore the question we are asking here is what is the expected value of the supremum of the canonical Gaussian process over some set T.

THEOREM 5.2.1. *Suppose that A is $n \times d$ $(n < d)$ has i.i.d. $\mathcal{N}(0,1)$ entries, then any solution $\hat{x}$ to $*$ satisfies*

$$\mathbb{E}\left[||\hat{x}(*) - x_0||_2\right] \leq \frac{Cw(\mathcal{K})}{\sqrt{n}},$$

where C is constant.

Observations. We have a few observations regarding the Gaussian width (Definition 5.1).

(1) If $A, B \subseteq \mathbb{R}^d$, then $w(A + B) = w(A) + w(B)$.
(2) If $a \in \mathbb{R}$, then $w(aA) = |a| w(A)$.
(3) If $A \subseteq \mathbb{R}^d$, then $w(A) = \frac{1}{2} w(A - A)$.

The last observation is helpful for us because it helps us picture what Gaussian width is telling us. Unless A is a degenerate set, this would be

$$\begin{aligned} w(A - A) &= \mathbb{E} \sup_{x,y \in A} \langle g, x - y \rangle \\ &= \mathbb{E} \sup_{x,y \in A} \langle \odot, x - y \rangle \cdot ||g||, \end{aligned}$$

where $\odot$ denotes a uniform random vector over the sphere (S^{d-1}). Note that $||g||$ and $\odot$ are the spherical and radial components of the Gaussian vector and are independent. Therefore

$$\begin{aligned} w(A - A) &= \mathbb{E} \sup_{x,y \in A} \langle \odot, x - y \rangle \cdot ||g|| \\ &\approx \sqrt{d} w_s(A), \end{aligned}$$

where $w_s(A)$ is the Spherical width of A.

But what is the Spherical Width $w_s(A)$? Look at it this way: take a random direction θ and measure the width of your set width respect to that direction (see Figure 5.1). In the end we average over the choice of the random direction.

DEFINITION 5.3 (Spherical Width). The Spherical Width of $T \subseteq \mathbb{R}^d$ is

$$w_s(T) := \mathbb{E}[\sup_{v \in T} \langle \odot, v \rangle], \quad \odot \sim \mathcal{U}(S^{d-1}).$$

In the end, note that going from the Gaussian Width to the Spherical Width is just multiplying by a constant $1/\mathbb{E}[||g||]$.

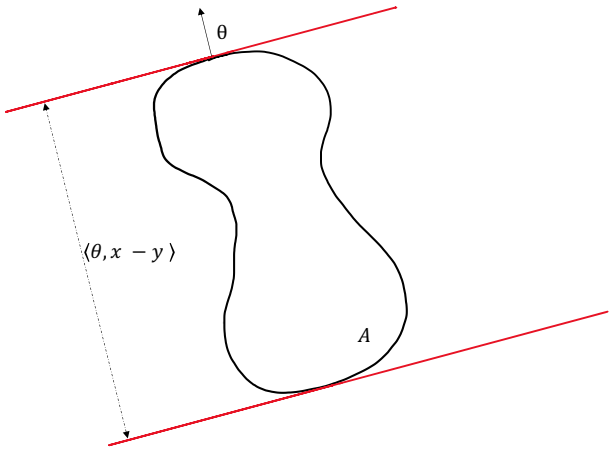


FIGURE 5.1. The width of A with respect to vector a random direction θ.

EXAMPLE 5.4 (unit ℓ_2 ball). What is the Gaussian Width of the unit ball?

$$w(B_2^d) = w(S^{d-1}) = \mathbb{E}||g|| = \sqrt{d}.$$

Therefore if our convex body is the unit ball i.e. $\mathcal{K} = B_2^d$, Theorem 5.2.1 implies that

$$\mathbb{E}||\hat{x}(*) - x_0||_2 \leq \frac{Cw(\mathcal{K})}{\sqrt{n}} \lesssim \frac{\sqrt{d}}{\sqrt{n}},$$

where by $\lesssim$, we mean that the inequality is correct up to some constant. Therefore, if all we knew was that our x_0 is in the unit ℓ_2 ball then our bound would not be great, as we would require $n \sim d$ samples to obtain constant error.

EXAMPLE 5.5 (unit ℓ_∞ ball). What is the Gaussian Width of the unit ℓ_∞ ball?

$$w(B_\infty^d) = \mathbb{E} \sup_{v \in B_\infty^d} \langle s, v \rangle = \mathbb{E}||g||_1 = cd$$

Therefore, the bound Theorem 5.2.1 would imply is

$$\mathbb{E}||\hat{x}(*) - x_0||_2 \leq \frac{Cw(\mathcal{K})}{\sqrt{n}} =\lesssim \frac{d}{\sqrt{n}}.$$

Therefore, just knowing that the entries are bounded does not give us an interesting bound as we would require $n \sim d^2$ samples to obtain constant error.

This reminds us of the first lectures where we looked at the uniform measure on a box and asked the question: how big is the ball inscribed inside this box? We saw that if we draw the ℓ_2 ball inside the ℓ_∞ ball, the mass we capture is basically nothing. Therefore Example 5.4 and Example 5.5 motivate us to look into ℓ_p balls, where $p \leq 2$.

EXAMPLE 5.6 (unit ℓ_1 ball). What is the Gaussian Width of the unit ℓ_1 ball?

$$w(B_1^d) = \mathbb{E} \sup_{v \in B_1^d} \langle s, v \rangle = \mathbb{E}||g||_\infty = c\sqrt{\log d}$$

Therefore, the bound Theorem 5.2.1 implies is

$$\mathbb{E}||\hat{x}(*) - x_0||_2 \leq \frac{Cw(\mathcal{K})}{\sqrt{n}} =\lesssim \sqrt{\frac{\log d}{n}}.$$

Therefore in this case we would require only $n \sim \log d$ [2] samples in order to obtain constant error, which is great.

6. Gaussian Complexity

DEFINITION 6.1 (Gaussian Complexity). The Gaussian Width of $T \subseteq \mathbb{R}^d$ is

$$\gamma(T) := \mathbb{E}\left[\sup_{v\in T}|\langle g, v\rangle|\right], \quad g \sim \mathcal{N}(0, I_d)$$

As you can see, Gaussian Width (Definition 5.1) seems to be related to Gaussian Complexity as they differ only in an absolute value.

Observations.

(1) In general,
$$w(T) \leq \gamma(T).$$
(2) If T is symmetric with respect to the origin i.e. $T = -T$, then
$$w(T) = \gamma(T).$$
(3) In general, the equality does not hold, but there exists some universal constant $C > 0$ such that for $\forall y \in T$
$$[w(T) + ||y||_2] \leq \gamma(T) \leq C\left[w(T) + ||y||_2\right].$$
(4) Note that $A - A$ is symmetric, therefore observation 2 of Gaussian Complexity 6.1, and observation 3 of Gaussian Width 5.1 imply that
$$w(A) = \frac{1}{2}w(A - A) = \frac{1}{2}\gamma(A - A).$$

7. Dudley's Bound

THEOREM 7.0.1 (Dudley's Bound). *Suppose $(X_t)_{t\in T}$ is a centered Gaussian process. Then*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \int_0^\infty \sqrt{\log N(T, d, \epsilon)}\, d\epsilon,$$

where the distance over T is

(Intrinsic Metric) $$d(s, t) = \mathbb{E}[X_s - X_t]^{1/2} = ||X_s - X_t||_2,$$

and $N(T, d, \epsilon)$ is the size (cardinality) of the smallest ϵ-net of T.

REMARK 7.1. If $(X_t)_{t\in T}$, is the Isonormal Gaussian process, then

$$d(s, t) = ||\langle g, s\rangle - \langle g, v\rangle||_{L^2} = ||s - v||_2.$$

Therefore Isonormal Gaussian processes have the property that for them, the intrinsic metric is equal to the metric of the ambient.

[2] This reminds us of the Johnson–Lindenstrauss lemma.

8. $M^\star$ Bound

The key point to the proof of Theorem 5.2.1 is to use the $M^\star$ bound.

THEOREM 8.0.1 ($M^\star$ bound). *Let $\mathcal{G}_{d,k}$ be the Grassmannian manifold of k-dimensional subspaces of $\mathbb{R}^d$. Note that $\mathcal{G}_{n,k}$ is compact. Suppose $T \subseteq \mathbb{R}^d$, $V \sim \mathcal{U}(\mathcal{G}_{d,k})$, then*

$$\mathbb{E}\left[\operatorname{diam}(T \cap V)\right] \leq \frac{Cw(T)}{\sqrt{k}}.$$

In the proof of Theorem 5.2.1 we think of the subspace $V \sim \mathcal{U}(\mathcal{G}_{d,k})$ as the $\ker A$, where A is a $k \times d$ matrix.

THEOREM 8.0.2. *Let $T \subseteq \mathbb{R}^d$, A be an $n \times d$ ($n < d$) matrix which has i.i.d. $\mathcal{N}(0,1)$ entries. Then*

$$\mathbb{E}\left[\operatorname{diam}(T \cap \ker A)\right] \leq \frac{Cw(T)}{\sqrt{k}},$$

for a constant C. Furthermore ,

$$\mathbb{E}\left[\sup_{x \in T} \operatorname{diam}(T \cap (x + \ker A))\right] \leq \frac{Cw(T)}{\sqrt{k}},$$

for a constant C.

9. Proof of Theorem 5.2.1

PROOF. Suppose $x_0, \hat{x} \in \mathcal{K}$, where $\hat{x}$ is our potential solution and x_0 is the unknown parameter. Then

$$y = A\hat{x} = Ax_0 \implies \hat{x} = x_0 + \ker(A).$$

We intend to bound $\mathbb{E}||\hat{x} - x_0||$. We may write

$$\begin{aligned} ||\hat{x} - x_0|| &\leq \operatorname{diam}(\mathcal{K} \cap (x_0 + \ker A)) \\ &\leq \max_{x' \in \mathcal{K}} \operatorname{diam}(\mathcal{K} \cap (x' + \ker A)). \end{aligned}$$

Taking the expectation from both sides and applying Theorem 8.0.2 finishes the proof

$$\begin{aligned} \mathbb{E}\left[||\hat{x} - x_0||\right] &\leq \mathbb{E}\left[\operatorname{diam}(\mathcal{K} \cap (x_0 + \ker A))\right] \\ &\leq \mathbb{E}\left[\max_{x' \in \mathcal{K}} \operatorname{diam}(\mathcal{K} \cap (x' + \ker A))\right] \\ &\leq \frac{Cw(\mathcal{K})}{\sqrt{n}}. \end{aligned}$$

□

CHAPTER 11

Lec 15: Mathematical Aspects of Neural Networks
Scribe: L. Zheng

1. Introduction

Let's start with a question of classification (supervised learning task). Given $\{(y_\ell, x_\ell)\}_{\ell=1}^M$, $y_\ell = \{0, 1\}$ are the binary labels and x_ℓ are the input data. We assume an unknown function with unknown parameter Θ, and the goal is to learn the parameter

$$y_\ell \approx f(x_\ell, \Theta)$$

In terms of Neural Networks, e.g. $1-$layer network,

$$f(x, \Theta) = \sigma(W^T x + b), \quad \Theta = (W, b)$$

where the $\sigma(\cdot)$ is the activation function/inverse link function, W is feature matrix and b is a shift term. The activation is chosen depends on your problem, e.g.

$$\sigma(x) = x$$
$$\sigma(x) = \log(1 + e^{-x})$$
$$\sigma(x) = \tanh(x)$$

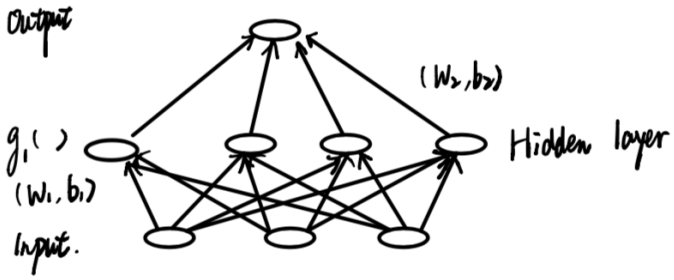
In multi-layer network, we then assume $f(x, \Theta) = g_k(\cdot, \Theta_k) \circ g_{k-1}(\cdot, \Theta_{k-1}) \circ \cdots \circ g_1(x, \Theta_1)$.

In the $2-$layer/1 hidden layer setting,

$$f(x) = g_2(W_2^T g_1(W_1^T x + b_1) + b_2)$$

note that the output is a scalar and g_1 are component-wise.

FIGURE 1.1. 1 hidden layer network.



The question is how to choose $\Theta = ((W_1, b_1), (W_2, b_2))$? We have some algorithms including

(1) Empirical risk minimization / M-estimation with Gradient descent
(2) Stochastic approximation / Stochastic gradient descent
(3) ...

The agenda for the following lectures:

(1) Approximation theory
(2) Shallow networks (the number of layer is fixed but the width of the layer to be infinite), Neural Tangent kernel (NTK), Random Features Regression, and double descent phenomenon, ...

2. Statistical Learning Theory

Given $(X_\ell)_{\ell=1}^M \sim \mathbb{P}_{\Theta_0}$ with unknown $\Theta_0 \in \Omega$ and loss function $L : \mathbb{R}^D \times \Omega \to \mathbb{R}$, the goal is to infer Θ_0, and we want to minimize the population risk

$$\Phi(\Theta) = \mathbb{E}_{X\sim\mathbb{P}_{\Theta_0}} L(X, \Theta)$$

where $\mathbb{P}_{\Theta_0}$ is the population.

In practice, we can minimize

(1) Empirical Risk of the entire dateset

$$\hat{R}(\Theta) = \frac{1}{M}\sum_{\ell=1}^{M} L(X_\ell, \Theta)$$

with gradient descent/global optimization, since the fluctuation $\hat{\varepsilon} = \hat{R} - \Phi$ is small
(2) Stochastic Approximation, i.e. iteratively optimize the loss over the chunk

$$L(X_\ell, \Theta) = \Phi + \hat{\varepsilon}_\ell$$

Next, the question is what can NN's learn?

In the setting of $1-$layer network, suppose $\sigma(x) = \log(1 + e^{-x})$, the loss function $L : [0, 1]^2 \to \mathbb{R}$ is $L(y, \sigma) = -(y \log \sigma + (1 - y) \log(1 - \sigma))$, and $g(x, \Theta) = \sigma(W^T x + b)$ where $\Theta = (W, b)$. So the empirical risk minimization

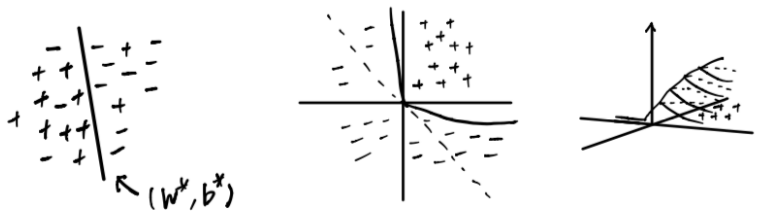
$$\begin{aligned}\Rightarrow \frac{1}{M}\sum_{\ell=1}^{M} L(y_\ell, x_\ell, \Theta) &= -\frac{1}{M}\sum_{\ell=1}^{M} y_\ell \log \sigma(W^T x_\ell + b) + (1 - y_\ell)\log(1 - \sigma(W^T x_\ell + b)) \\ &= -\frac{1}{M}\sum_{\ell=1}^{M} y_\ell \log \hat{y}_\ell + (1 - y_\ell)\log(1 - \hat{y}_\ell)\end{aligned}$$

We denote $\hat{y}_\ell = \sigma(W^T x_\ell + b)$, and the problem is exactly the logistic regression! This is good for the linear boundary, but what if we have nonlinear boundary?

Then, consider a $2-$layer network

$$\sigma(\sigma(e_1 \cdot x) + \sigma(e_2 \cdot x) - 2\log 2)$$

FIGURE 2.1. 1−layer network for linear boundary (left), 2−layer network for nonlinear boundary in 2−dimension (middle) and 3−dimension (right).



3. Approximation Theory for NNs

$$\mathcal{F} = \{\underline{a}, \underline{b}, \underline{W} : \sum_i a_i \sigma(w_i^T x + b_i))\}$$

this is a 2−layer network: $g_2(\sigma, \Theta) \cdot g_1(\sigma, \underline{a}) = a \cdot \sigma$, what can we learn if the activation function in the final layer is linear.

Theorem 3.1. *(Universal Approximation (Cybenko '89)) Suppose $\sigma : \mathbb{R} \to [0, 1]$ is continuous monotone, and*

$$\sigma(t) \to \begin{cases} 1 & t \to \infty \\ 0 & t \to -\infty \end{cases}$$

$\mathcal{F}$ is dense in $(C([0, 1]^n, [0, 1]), ||\cdot||)$, i.e. for $G \in C$ there exist $f_p \in \mathcal{F}$, such that

$$\sup_x |G(x) - f_p(x)| < \epsilon$$

4. Algebraic Complexity

$$f : \{0, 1\}^n \to \{0, 1\}$$

The next question is how many terms (or how wide) do we need to approximate the function, i.e. Depth v.s. Width.

Theorem 4.1. *For any $d \geq 1$, there exist $\mathbb{P} \in \mu(\mathbb{R}^d)$, $g : \mathbb{R}^d \to \mathbb{R}$ with $Supp(g) \subseteq B_{\sqrt{d}}$ such that g is expressible, i.e. can be approximated by a polynomial ($\sim d^5$) width 3−layer network but such that any function $f \in \mathcal{F}_{2-layer}$ with width at most Ce^{cd}*

$$||g - f||_{L^2(\mathbb{P})} > c$$

CHAPTER 12

Lec 16: Kernels, Networks, Overparameterization
Scribe: L. Zheng

Last time we introduce the universal approximation theorem

$$\mathcal{F} = \{\underline{a}, \underline{b}, \underline{W} : \sum_i a_i \sigma(w_i^T x + b_i))\}$$

is dense in $(C([0,1]^n, [0,1]), ||\cdot||)$. Today, we discuss the "Shallow" or "Wide" networks, including

(1) Overparamerization + "Double Descent"
(2) Mean field methods for Wide Neural Networks

1. Overparameterization + Double descent

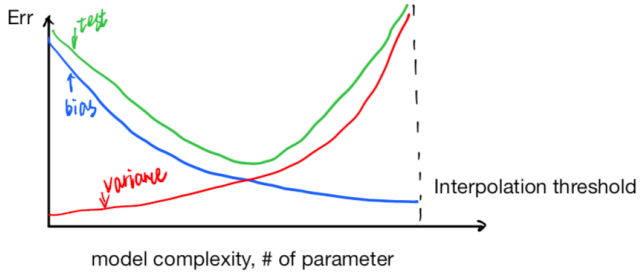
The goal for today is to understand how well can we do regression using $\mathcal{F}$ for

$$y_\ell = f(x_\ell) + \epsilon_\ell$$

The well-known Bias-Variance Trade-off for test error is

$$\mathbb{E}((f - \hat{f})^2) = \underbrace{\mathbb{E}((f - \mathbb{E}\hat{f})^2)}_{\mathbb{E}\text{Bias}^2} + Var(\hat{f})$$

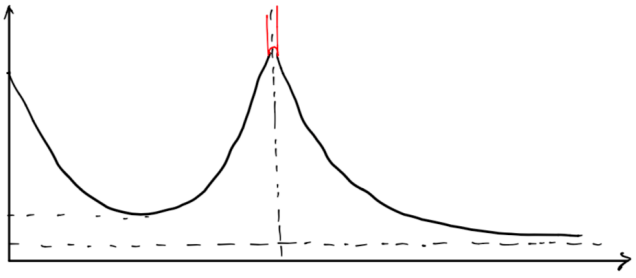
FIGURE 1.1. Bias-Variance Trade-off for test error.



Modern networks tend to be overparameter, but there is a double descent phenomenon

(1) Variance could decrease after the "interpolation" threshold
(2) Bias could be non-monotone near "interpolation"

FIGURE 1.2. Double Descent.



2. Random Feature Models

Given $\{(y_\ell, x_\ell)\}_{\ell=1}^M$, $x_\ell \sim \text{Unif}(S^{d-1}(\sqrt{d}))$ a sphere with $r = \sqrt{d}$, and

$$y_\ell = f_*(x_\ell) + \epsilon_\ell$$

where $\epsilon_\ell \sim \text{IID}(0, 1)$. The goal is to infer f_*. There are many methods and one popular class is the kernel methods:

$$\min R((y_\ell, x_\ell), f) + \Omega(||f||_{\mathcal{H}}^2) \qquad (*)$$

these functions are called "Reproducing Kernel Hilbert Space for K". ($f \approx \sum \alpha_i \sigma(W_i^T x)$ which reminds us the kernel methods with random weights.)

The idea of Neural Tangent methods is to focus on the evolution of function

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial t}(\sum L((y_\ell, x_\ell), f_\theta)) = \sum \frac{\partial L}{\partial f}\Big|_\ell \frac{\partial f(x_\ell)}{\partial \theta}\frac{\partial \theta}{\partial t}$$

where $L((y, x), f)) = (y - f(x))^2$

Theorem 2.1. *The minimizer of* $(*)$ *is*

$$f = \frac{1}{M}\sum_{i=1}^M \alpha_i K(x_i, x)$$

with

$$K(x, y) = \mathbb{E}_z\left[\sigma\left(\frac{\langle x, z\rangle}{\sqrt{d}}\right)\sigma\left(\frac{\langle z, y\rangle}{\sqrt{d}}\right)\right]$$

where $z \sim \text{Unif}(S^{d-1}(\sqrt{d}))$ However, the computation in high-dimension is very expensive, and let's do the Monte Carlo.

2.1. Random Features Ridge Regression. Suppose $(z_i)_{i=1}^N \sim \text{Unif}(S^{d-1}(\sqrt{d}))$ independently and

$$K_N(x, y) = \frac{1}{N}\sum \sigma\left(\frac{\langle x, z_i\rangle}{\sqrt{d}}\right)\sigma\left(\frac{\langle z_i, y\rangle}{\sqrt{d}}\right)$$

then, by law of large number $K_N \to K$.

Sampling uniformly from from the sphere is fast, and thus compute K_N is faster than the integral K. The idea of random features regression is to do the penalized least square

regression by using kernel methods K_N.

$$\min_a \frac{1}{m}\sum (y_i - \sum_{j=1}^{N} a_j \sigma\left(\frac{\langle x_i, z_j\rangle}{\sqrt{d}}\right))^2 + \lambda||a||_2^2$$

We impose the ridge penalty to avoid complex model for this type of regression task. We obtain the minimizer $\hat{a}(\lambda)$ and then $\hat{f}(x, \hat{a}, \Theta)$ where $\Theta = (z_1, ..., z_N)$

Then, we focus on the test error

$$R(f_*, x, \Theta, \lambda) = \mathbb{E}[(f_*(x) - f(x, a, \Theta))^2] + C$$

for some $x \sim \text{Unif}(S^{d-1}(\sqrt{d}))$ and constant C independently from the noise ϵ. We can compute this R explicitly in "interesting" regime as follows

$$\psi_1 = \frac{N}{d}, \psi_2 = \frac{M}{d}, d \to \infty, \psi_1, \psi_2 > 0$$

[**3**] assume that

$$f_*(x) = \overbrace{\beta_{d,0} + \langle \beta_{d,1}, x\rangle}^{\text{Linear part}} + f_d^{NL}(x)$$

and f^{NL} is a Gaussian Process with $\mathbb{E}f^{NL}(x)f^{NL}(y) = \Sigma(\frac{\langle x,y\rangle}{d})$

Aside: (Schoenberg '40)

$$f^{NL}(x) = \sum \alpha_p f_p(x) \qquad f_p(x) = \sum g_{i1,...,ip} x_{i1}...x_{ip}$$

where $g_{i1,...,ip}$ IID $N(0,1)$. A toy example: $p = 2$, $\Sigma(t) = t^2$ then $f^{NL} = \langle x, Gx\rangle$ where $G \sim GOE$.

Theorem 3.1. *Suppose $X = [x_1, ...x_M]^{M\times d}$, where $x_\ell \sim$ Unif$(S^{d-1}(\sqrt{d}))$, $\Theta = (\theta_1, ..., \theta_n) \sim$ Unif$(S^{d-1}(\sqrt{d}))$, $\hat{\sigma}(0), \hat{\sigma}(1) < \infty$ and $||\sigma||_{L^2}^2 - \hat{\sigma}(0)^2 - \hat{\sigma}(1)^2 < \infty$*

$$R(f_*, x, \Theta, \lambda) \xrightarrow{L^1} F_1 B(\psi_1, \psi_2, \hat{\sigma}, \lambda) + F_2 V + F_*^2$$

where F_1, F_2 are constant, B is the bias, V is the limit of variance, and $F_*^2 = \Sigma(1)$ is the correction term. We get double descent.

The random features regression in this regime can only learn the linear part, but even in this linear part, double descent appears!

Bibliography

[1] Bordenave, C., Caputo, P. and Chafaï, D., 2011. Spectrum of non-Hermitian heavy tailed random matrices. Communications in mathematical physics, 307(2), pp.513-560.
[2] Anderson, Guionnet, Zeitouni, Random Matrix Theory
[3] Mei, S. and Montanari, A., The generalization error of random features regression: Precise asymptotics and double descent curve, arXiv:1908.05355